Caracterização e Detecção de Sequestros de Prefixo na Internet

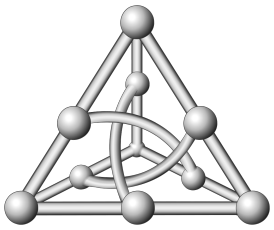
Adriano Bastos de Carvalho

Dissertação de Mestrado

Área de Concentração: Redes de Computadores

Orientação: Prof. Ronaldo Alves Ferreira, Ph.D.

Coorientação: Prof. Carlos Alberto da Silva, Ph.D.



Faculdade de Computação
Universidade Federal de Mato Grosso do Sul
Fevereiro de 2025

Caracterização e Detecção de Sequestros de Prefixo na Internet

Adriano Bastos de Carvalho

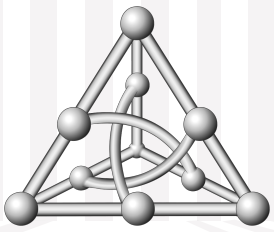
Dissertação de Mestrado

Área de Concentração: Redes de Computadores

Orientação: Prof. Ronaldo Alves Ferreira, Ph.D.

Coorientação: Prof. Carlos Alberto da Silva, Ph.D.

Apresentado em cumprimento parcial dos requisitos para o grau de Mestre em Ciência da Computação



Faculdade de Computação
Universidade Federal de Mato Grosso do Sul
Fevereiro de 2025

Agradecimentos

Ao nosso Criador, Deus Pai, e Jesus por ter nos permitido chegar até este momento com a capacidade e conhecimento necessário para realização deste curso.

A minha amada esposa, Valéria, e aos meus filhos, Bruno e Diego, pela compreensão e apoio durante a realização deste curso e também durante todo o percurso percorrido até este momento. Graças a eles tive motivação e condições de me dedicar o máximo possível a pesquisa que resultou neste trabalho.

Aos meus pais, Vicente e Alzira, que me não mediram esforços para permitir que seus filhos recebessem a melhor educação possível, e aos meus irmãos, Luciano e Matheus, que são exemplos de dedicação ao que fazem.

A todos os integrantes do 6º Centro de Telemática de Área e a todos os amigos que me apoiaram durante a realização do curso, em especial ao amigo Coronel Cláudio, pois sem seu apoio não teria conseguido a autorização para realizar este curso.

Aos professores Pedro (FURG) e Fabrício (UEMS) pela orientação na minha qualificação e pelo trabalho realizado em conjunto. Ao professor Brivaldo (UFMS) por todo auxílio desde o início do curso nas conduções das pesquisas e por todo conhecimento compartilhado.

Ao professor Carlos Alberto (UFMS), além de me acompanhar durante todo o curso como coorientador, me ajudou com várias propostas de tema de pesquisa que poderiam ser de interesse do Exército para que eu pudesse solicitar a autorização para realizar o curso.

Ao professor Ronaldo (UFMS), pois graças as suas orientações e correções este trabalho foi possível. Sem a sua orientação, as ideias inicialmente vagas que eu tinha no começo do curso não poderiam ser aprofundadas e se tornarem o trabalho aqui apresentado, o artigo que foi publicado no SBSeg 2024 e o artigo submetido ao SBRC 2025.

A todos os professores da Facom/UFMS por terem compartilhado seus conhecimentos comigo e a toda equipe da secretaria pelos esclarecimentos e apoio prestado.

Resumo

O protocolo de roteamento BGP (*Border Gateway Protocol*) não possui mecanismos nativos de segurança, permitindo que atores maliciosos manipulem os anúncios de rota ou anunciem prefixos que não lhe pertencem. Quando um sistema autônomo (AS – *Autonomous System*) anuncia um prefixo que não lhe pertence, ocorre um sequestro de prefixo, o que pode deixar o AS legítimo inacessível, desviar o tráfego para roubo de informações ou permitir a utilização indevida dos endereços sequestrados (*e.g.,* para envio de *spam*). Alguns trabalhos propõem soluções para esse problema, como RPKI, BGPSec e ASPA, mas essas soluções ainda não foram amplamente implementadas para eliminar o problema.

A primeira parte deste trabalho utiliza um conjunto extensivo de simulações, com dados reais, para caracterizar a vulnerabilidade a sequestros de prefixo de 29 redes militares, revelando que redes mais conectadas e com vizinhos distribuídos geograficamente são menos afetadas. O estudo realizado também discute possibilidades para tornar os sistemas de roteamento dessas redes mais robusto.

Trabalhos recentes utilizam aprendizado de máquina para identificar esses sequestros, mas os modelos são complexos e do tipo caixa-preta, tornando inviável determinar se utilizam as *features* mais adequadas. A segunda parte deste trabalho aplica técnicas de Inteligência Artificial Explicável (XAI) para avaliar e melhorar um modelo de detecção de sequestros de prefixo proposto recentemente. A partir de uma análise do modelo original com 28 *features*, foram criados dois modelos reduzidos com 11 e 5 *features*, que produzem resultados sem diferenças estatísticas do modelo completo, mas reduzem o tempo de processamento em mais de 31% (9 min por dia) e o espaço de armazenamento total necessário em mais de 36% (970 MB em 160 dias). Quando os resultados obtidos pelos modelos reduzidos com base em novos enlaces identificados são avaliados, o modelo de 5 *features* se mostrou mais preciso em 0,1152 em relação ao modelo original, demonstrando a importância da correta seleção de *features*. Analisando os sequestros simulados de redes militares, até 77% dos ataques podem passar despercebidos, mesmo com a melhor ferramenta disponível para detectar sequestros com origem forjada. Além da redução de *features*, duas abordagens que buscam melhorar o modelo também são apresentadas, uma verificando o impacto no modelo caso seja possível a obtenção de novas informações para incrementar os valores obtidos para uma das *features* de bidirecionalidade e outra verificando o resultado do modelo com uma nova amostragem para o treinamento. A primeira abordagem resultou em aumento no F1-*score* para ambas as classes e a segunda em aumento no MCC (*Matthews Correlation Coefficient*) do modelo de -0,0530 para 0,3165.

Palavras-chave: *Protocolo BGP, Segurança, Sequestro de prefixo, Inteligência Artificial Explicável.*

Abstract

The Border Gateway Protocol (BGP) lacks native security mechanisms, allowing malicious actors to manipulate route announcements or advertise prefixes they do not own. When an Autonomous System (AS) advertises a prefix it does not own, a prefix hijack occurs, which can render the legitimate AS unreachable, redirect traffic to steal information, or enable the misuse of the hijacked addresses (e.g., for sending spam). Some works propose solutions to this problem, such as RPKI, BGPSec, and ASPA, but these solutions have not yet been widely implemented to eliminate the issue.

The first part of this work uses an extensive set of simulations with real data to characterize the vulnerability of 29 military networks to prefix hijacks, revealing that networks with higher connectivity and geographically distributed neighbors are less affected. The study also discusses possibilities for making the routing systems of these networks more robust.

Recent research has employed machine learning to identify these hijacks, but the models are often black boxes and complex, making it challenging to determine whether they use the most appropriate features. The second part of this work applies eXplainable Artificial Intelligence (XAI) techniques to evaluate and improve a recently proposed prefix hijack detection model. From an analysis of the original model with 28 features, two reduced models were created with 11 and 5 features, respectively. These reduced models produce results with no statistical difference from the complete model but reduce processing time by over 31% (9 min per day) and total storage space required by more than 36% (970 MB in 160 days). When the results obtained by the reduced models using new links identified are evaluated, the 5-feature model proved to be 0.1152 more accurate than the original model, demonstrating the importance of proper feature selection. Evaluating the simulated hijacks of military networks, up to 77% of attacks may go undetected, even with the best available tool for detecting hijacks with forged origins. In addition to feature reduction, two approaches to improving the model are also presented: one assesses the impact on the model if it were possible to obtain new information to enhance the value of the bidirectionality feature, and the other examines the model's results with a new training dataset. The first approach resulted in an increase in the F1-score for both classes, while the second improved the Matthews Correlation Coefficient (MCC) from -0.0530 to 0.3165.

Keywords: *BGP Protocol, Security, Prefix Hijacking, eXplainable Artificial Intelligence.*

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Siglas

AM	Aprendizado de Máquina
AS	*Autonomous System*
ASN	*Autonomous System Number*
ASPA	*Autonomous System Provider Authorization*
BGP	*Border Gateway Protocol*
BGPsec	*Border Gateway Protocol Security*
c2p	*Customer-to-Provider*
CAIDA	*Center for Applied Internet Data Analysis*
DDoS	*Distributed Denial-of-Service*
DFOH	*Detect Forged-Origin Hijacks*
DT	*Decision Tree*
FN	Falsos Negativos
FP	Falsos Positivos
IA	Inteligência Artificial
IANA	*Internet Assigned Numbers Authority*
IC	Intervalo de Confiança
ICMP	*Internet Control Message Protocol*
IP	*Internet Protocol*
IPv4	*Internet Protocol* versão 4
IPv6	*Internet Protocol* versão 6
ISP	*Internet Service Provider*

IXP	*Internet eXchange Point*
LDC	*Load Distribution Change*
LLM	*Large-Language Models*
LSTM	*Long Short Term Memory*
MANRS	*Mutually Agreed Norms for Routing Security*
MCC	*Matthews Correlation Coefficient*
MOAS	*Multiple Origin Autonomous System*
p2c	*Provider-to-Customer*
p2p	*Peer-to-Peer*
PHAS	*Prefix Hijack Alert System*
PTT	Ponto de Troca de Tráfego
RFC	*Request for Comments*
RIB	*Routing Information Base*
RIR	*Regional Internet Registry*
ROA	*Route Origin Authorization*
ROV	*Route Origin Validation*
RPKI	*Resource Public Key Infrastructure*
RTT	*Round Trip Time*
s2s	*Sibiling-to-Sibiling*
TFN	Taxa de Falsos Negativos
TFP	Taxa de Falsos Positivos
TVN	Taxa de Verdadeiros Negativos
TVP	Taxa de Verdadeiros Positivos
VN	Verdadeiros Negativos
VP	Verdadeiros Positivos
XAI	*eXplainable Artificial Intelligence*

Capítulo 1

Introdução

Surgida na década de 1960 em um ambiente confiável e com acesso restrito, a Internet tinha como objetivo inicial conectar centros de pesquisa e instituições militares nos Estados Unidos. Entretanto, a partir do final da década de 1980, seu uso se expandiu para os setores privado e comercial, resultando em um crescimento significativo e consequente globalização [79]. Com essa globalização, a Internet passou a abrigar usuários de diversas naturezas, incluindo atores maliciosos, como evidenciado pelos crimes cibernéticos frequentes.

A Internet é formada por redes administradas de forma independente que se conectam para prover conectividade aos usuários. Uma rede, ou um conjunto de redes, administrada por uma determinada entidade é denominada de Sistema Autônomo (AS – *Autonomous System*). Um AS define suas próprias políticas de roteamento e acordos com outras redes visando implementar um serviço de entrega de pacotes fim-a-fim. Cada AS possui um conjunto de endereços IP, que é representado por um ou mais prefixos de rede, e um número que o identifica denominado ASN (*Autonomous System Number*).

Os sistemas autônomos na Internet trocam informações de roteamento utilizando o protocolo BGP (*Border Gateway Protocol*) [67] que oferece vários mecanismos para suportar políticas complexas de roteamento. A construção de uma rota em BGP começa quando um *AS origem* anuncia um prefixo IP aos seus ASes vizinhos. As rotas são então propagadas por meio de mensagens de atualização BGP entre os ASes. O *AS-path* é a sequência de ASes atravessados pela rota até atingir o AS origem. O *AS-path* é empregado pelo BGP para prevenir *loops* e também no processo de seleção da melhor rota para um prefixo de destino. BGP escolhe a melhor rota para um destino (*i.e.,* prefixo) utilizando uma sequência de critérios que inclui o atributo de preferência local (*LocalPref*), origem da rota, comprimento de *AS-path*, etc. [67].

Apesar de toda sua flexibilidade, BGP não implementa nativamente técnicas de validação e autenticação de anúncios de rota recebidos. A ausência de mecanismos de segurança em BGP permite que um AS anuncie prefixos de outros ASes e altere os anúncios de rota antes de repassá-los, incluindo modificações no *AS-path*. Alterações maliciosas no *AS-path* podem ocultar a origem ilegítima de um prefixo e *sequestrar* o tráfego destinado a ele [23]. Um sequestro pode resultar na indisponibilidade dos recursos da rede do prefixo afetado, na interceptação ou alteração da comunicação e no

uso dos endereços para atividades maliciosas. Os sequestros podem ser acidentais, devido a erros de configuração (*i.e., vazamento de rota*), ou maliciosos, com intenções específicas.

Sequestros de prefixo não são amplamente divulgados pelas vítimas devido a possíveis impactos negativos na reputação de suas instituições. Entretanto, alguns casos públicos demonstram a gravidade do problema. Por exemplo, em 24 de fevereiro de 2008, o AS 17557 da Pakistan Telecom anunciou o prefixo 208.65.153.0/24 para bloquear o acesso de seus clientes ao YouTube. Esse anúncio, mais específico que o do AS 36561 pertencente à Google, redirecionou (*i.e., sequestrou*) o tráfego para a rede da Pakistan Telecom, causando instabilidade e indisponibilidade de acesso ao YouTube por mais de duas horas [70]. Esse exemplo ilustra um sequestro de prefixo usado por um país para censurar conteúdo na Internet. Porém, técnicas de sequestro de prefixo estão se aprimorando e sendo usadas para crimes. Em 3 de fevereiro de 2022, por exemplo, um sequestro de prefixo permitiu que hackers roubassem cerca de 1,9 milhão de dólares em criptomoedas da plataforma KlaySwap [78]. Mais recentemente, em 27 de junho de 2024, um sequestro afetou o serviço de DNS da Cloudflare degradando seu desempenho e causando indisponibilidades [38]. O sequestro intencional de prefixos pode ter vários objetivos, como divulgar páginas maliciosas (*phishing*), enviar spams, roubar informações, falsificar certificados digitais, entre outros [7, 23, 81].

O número crescente de problemas de segurança com BGP [11, 46, 50], que tem causado indisponibilidades frequentes de partes da Internet, levou pesquisadores e operadores de rede a propor vários mecanismos para melhorar sua segurança. Entre eles estão RPKI (*Resource Public Key Infrastructure*) [11], que oferece uma infraestrutura de chaves públicas para autorização e validação da origem de rotas, BGPSec [46], que permite a assinatura de todos os anúncios de rota, MANRS (*Mutually Agreed Norms for Routing Security*) [33], que define um conjunto de ações que os ASes devem implementar, como filtragem e validação de rotas, para evitar sequestros de prefixo, vazamento de rotas e falsificação (*spoofing*) de endereços [31, 32], e mais recentemente ASPA (*Autonomous System Provider Authorization*) [4], que busca validar as conexões entre ASes com o uso de RPKI. Entretanto, essas propostas ainda enfrentam resistências e não são amplamente implementadas, deixando o sistema de roteamento da Internet vulnerável a ataques e falhas de configuração.

A primeira parte deste trabalho apresenta uma caracterização da vulnerabilidade a sequestros de prefixo de ASes utilizados pelas Forças Armadas de países do G20. Em cenários críticos, as Forças Armadas mantêm infraestruturas segregadas e isoladas da Internet. No entanto, para oferecer serviços ou interagir com seus membros ou com a população civil em geral, disponibilizam sistemas acessíveis pela Internet, como serviços de alistamento, gestão de informações e acesso interno via VPN para militares em missões externas. Essa exposição, contudo, torna-os potenciais alvos de ataques cibernéticos. Por exemplo, conflitos globais recentes, como a guerra Rússia-Ucrânia, as tensões entre Israel e Hamas e disputas no Oriente Médio, evidenciam essa ameaça. Recentemente, a Rússia conduziu ataques cibernéticos contra a Ucrânia antes de ações militares [80]. No conflito entre Israel e Hamas, ambos os lados enfrentam ofensivas cibernéticas de grupos de *hackers* ativistas [61], evidenciando um campo de batalha sem fronteiras claras. Assim, identificar

e mitigar vulnerabilidades nesses sistemas é essencial para assegurar a segurança e a soberania nacionais.

Este estudo de caso apresenta a primeira caracterização da vulnerabilidade da infraestrutura de roteamento das redes militares das principais economias globais contra ataques de sequestro de prefixo. Para isso, três abordagens foram empregadas. A primeira avalia a vulnerabilidade das redes militares contra ataques de sequestro de prefixo utilizando simulações com topologia da Internet obtida com dados de medição disponibilizados pela CAIDA. Já a segunda consiste na análise de dados públicos de roteamento, que também foram usados para a construção do simulador, para entender as atuais práticas de roteamento das redes militares e estimar suas vulnerabilidades. Ao longo dessas etapas, foram analisados dados coletados em 530 pontos de observação entre os meses de fevereiro e abril de 2024, que totalizam mais de 735 milhões de rotas por dia. Por fim, a terceira avalia a efetividade dos recursos existentes para detecção de sequestros em situações de ataques realistas a redes militares.

Por ser uma infraestrutura crítica para a segurança e soberania de um país, a caracterização da vulnerabilidade de uma rede militar a sequestros de prefixo contribui em várias dimensões: *(i)* compreensão da gravidade do problema; *(ii)* identificação de padrões e tendências; *(iii)* proposição de medidas de mitigação; *(iv)* avaliação da eficácia de mecanismos de segurança; *(v)* apoio à tomada de decisão; *(vi)* impacto na segurança e estabilidade da Internet.

A segunda parte deste trabalho foca na detecção do sequestro de prefixo. Como ainda não se pode eliminar sequestros de prefixo na Internet, vários trabalhos recentes buscam detectá-los para alertar os operadores tão logo eles sejam identificados, visando mitigar o problema [39, 43, 65, 72, 75]. Muitas dessas abordagens têm empregado aprendizado de máquina devido à abundância de dados de roteamento disponíveis em coletores públicos [60, 66] e informações sobre ASes disponíveis em bancos de dados de registros regionais (RIRs, *Regional Internet Registries*) [54], os quais são utilizados para treinar os modelos [39, 74, 81]. No entanto, os modelos são do tipo caixa-preta (*black box*) e estão se tornando cada vez mais complexos, com uma grande quantidade de *features* que consomem muito tempo para serem computadas e que nem sempre melhoram o desempenho dos modelos.

Sem uma análise detalhada do funcionamento de um modelo de aprendizado de máquina e de como suas inferências ocorrem, *features* de pouca ou nenhuma relevância podem ser utilizadas, resultando em desperdício de espaço de armazenamento e tempo de processamento para cálculo das *features*, treinamento do modelo e inferência. Alguns trabalhos recentes propõem o uso de Inteligência Artificial Explicável (XAI, do inglês *eXplainable Artificial Intelligence*) para se entender e explicar as decisões de um modelo caixa-preta [40, 68]. Trustee [40], por exemplo, gera um modelo interpretável na forma de uma árvore de decisão que captura e explica o comportamento de um modelo caixa-preta. Um operador de redes pode analisar a árvore de decisão para verificar se o modelo está tomando decisões coerentes com o conhecimento existente do domínio do problema e identificar quais *features* contribuem ou não para as inferências.

Este trabalho utiliza técnicas e ferramentas de XAI para avaliar e aperfeiçoar um modelo caixa-preta para detecção de sequestros de prefixo proposto recentemente e

considerado como o estado-da-arte para detecção de sequestros com alteração no *AS-path*. O modelo, utilizado no sistema DFOH (*Detects Forged Origin Hijacks*) [39], utiliza 28 *features* extraídas de anúncios BGP [60], dados de registros regionais [54], bases da CAIDA [12] e do PeeringDB [62]. Na análise de funcionamento do DFOH, os autores obtiveram uma taxa de verdadeiros positivos de 0,909 e de falsos positivos de 0,019, resultando numa precisão de 0,9795 [39].

Uma avaliação do modelo de DFOH com a ferramenta Trustee possibilitou a criação de dois modelos reduzidos, um com 11 e outro com apenas 5 das 28 *features* originais. Validações dos modelos reduzidos, utilizando os métodos descritos em [39], mostram que os modelos reduzidos produzem resultados de precisão e *recall* dentro do mesmo intervalo de confiança do modelo completo, não apresentando, portanto, diferenças estatisticamente significativas. Por exemplo, o modelo com 5 *features*, apesar de uma redução superior a 80% nas *features*, produz uma diferença na média do F1-Score de apenas 0,0001. O número reduzido de *features* nos novos modelos diminui os tempos de processamento para cálculo das *features*, treinamento e inferência em mais de 30% e 37%, correspondendo a uma redução de 9,7 e 11,6 minutos por dia de processamento, para os modelos com 11 e 5 *features*, respectivamente. O espaço necessário para armazenar as *features* também reduz em mais de 59% e 71%, correspondendo a uma redução de 82,3 MB e 97,8 MB para 160 dias de informação. Considerando a soma do espaço utilizado pelas *features* e do modelo utilizado para o seu cálculo, a redução fica em 36% (970,1 MB) e 68% (1823,7 MB) para os modelos com 11 e 5 *features*, respectivamente.

Utilizando os três modelos para inferir as classes de novos enlaces surgidos ao longo de um período de 40 dias, foram observadas diferenças de 6% e 7,8% entre o modelo original e os modelos reduzidos com 11 e 5 *features*, respectivamente. Como não há uma verdade absoluta (*ground truth*) para esses novos enlaces, aqueles que ainda foram observados nos coletores em períodos posteriores (superiores a um mês) foram considerados legítimos. A justificativa para essa classificação é que sequestros de prefixo são realizados por períodos curtos de tempo [81], ou seja, enlaces suspeitos, que podem indicar sequestros de prefixo, só são observados nos coletores por períodos curtos de tempo. Nessa avaliação, o modelo com 11 *features* gerou um número menor de falsos positivos ao identificar corretamente um número maior de enlaces legítimos. Por outro lado, o modelo com 5 *features* obteve precisão e *recall* superiores ao modelo original para os enlaces suspeitos. Esses resultados surpreendentes sugerem que o modelo original contém *features* que não contribuem na separação das classes e, em alguns casos, até prejudicam o desempenho do modelo. O Capítulo 6 apresenta uma discussão qualitativa das *features* utilizadas nos modelos e possíveis motivos para melhorias de desempenho dos modelos reduzidos. Uma vantagem adicional dos modelos reduzidos é que, por serem computacionalmente mais leves, eles podem ser retreinados com maior frequência e, assim, capturar mais rapidamente mudanças nas distribuições dos dados e armazenar os valores para um período maior de tempo por necessitarem de menos espaço.

Uma análise com técnica de explicabilidade local utilizando o Lime [68] foi realizada para validar os resultados obtidos com a ferramenta Trustee, apresentando resultados consistentes e corroborando a análise anterior. Como resultado das análises, duas abordagens foram investigadas para melhorar a detecção de sequestros: uma baseada no

incremento de informações sobre enlaces bidirecionais e outra utilizando uma amostragem alternativa para o treinamento do modelo.

Na primeira abordagem, com o uso de dados sintéticos, verificou-se que a identificação de mais enlaces bidirecionais aumentou o F1-*score* do modelo em ambas as classes. Na segunda, o treinamento foi realizado com novos enlaces observados em períodos anteriores, validados como suspeitos ou legítimos com base em informações de períodos posteriores. A métrica MCC (*Matthews Correlation Coefficient*) [22] foi utilizada para comparar os modelos. Com essa abordagem, o modelo de 5 *features* alcançou um MCC de 0,3429, enquanto o modelo original obteve apenas -0,0077.

1.1 Organização do Texto

Este trabalho está organizado da seguinte maneira. O Capítulo 2 apresenta os conceitos básicos necessários para o entendimento de sequestros de prefixo e o que os tornam possíveis. Esse capítulo apresenta conceitos sobre a organização da Internet em sistemas autônomos e seus relacionamentos mais comuns (Seção 2.1), protocolo BGP (Seção 2.2), coletores de informações BGP de acesso público (Seção 2.3), principais técnicas propostas para melhorar a segurança do protocolo BGP (Seção 2.4), algumas definições básicas sobre Inteligência Artificial Explicável (Seção 2.5) e as principais métricas utilizadas para se analisar o resultado de modelos de Aprendizado de Máquina (AM) (Seção 2.6).

O Capítulo 3 discute em detalhes sequestros de prefixo, incluindo suas principais motivações (Seção 3.1), classificações (Seção 3.2), e alcance (Seção 3.3). O Capítulo 4 revisa trabalhos relacionados, agrupados pelo foco: caracterização de Sistemas Autônomos maliciosos (Seção 4.1), identificação de sequestro de prefixos no plano de dados (Seção 4.2), no plano de controle (Seção 4.3), e em ambos os planos (Seção 4.4).

O Capítulo 5 foca na caracterização de sequestros de prefixos anunciados por redes de Forças Armadas de alguns países. O capítulo também descreve o simulador de sequestro desenvolvido para realizar análises sem impactar a Internet (Seção 5.1), as características dos ASes selecionados como vítimas (Seção 5.2), a capacidade de contaminação dos sequestros simulados (Seção 5.3) e a capacidade de observação pelos coletores (Seção 5.4). Além disso, o capítulo discute a resiliência dos ASes ao sequestro (Seção 5.5) e o impacto do uso de *prepend* na origem dos anúncios (Seção 5.6).

O Capítulo 6 apresenta a análise conduzida para aprimorar a detecção de sequestros de prefixo com origem forjada. Inicialmente, o capítulo descreve o modelo de referência, o DFOH [39] (Seção 6.1). Em seguida, aborda os resultados obtidos com a ferramenta Trustee (Seção 6.2), detalha o ambiente e os dados utilizados (Seção 6.3), apresenta os resultados dos novos modelos (Seção 6.4) e analisa o uso da ferramenta Lime (Seção 6.5). Na sequência, o capítulo analisa a capacidade dos modelos identificarem os sequestros simulados na caracterização (Seção 6.6) e discute as melhorias propostas para os modelos (Seção 6.7).

O trabalho é concluído no Capítulo 7, que apresenta as considerações finais e possibilidades para trabalhos futuros. Os trabalhos publicados e submetidos estão listados no Apêndice A.

Capítulo 2

Conceitos Básicos

Este capítulo apresenta alguns conceitos básicos essenciais para a compreensão do que é um sequestro de prefixo, suas causas e consequências. Ele inicia com a definição de sistemas autônomos, como eles se interconectam e as principais relações entre eles, que são conceitos fundamentais para se entender a estrutura da Internet (Seção 2.1). Em seguida, o capítulo descreve brevemente o protocolo BGP (*Border Gateway Protocol*), que é o protocolo de roteamento *de facto* da Internet (Seção 2.2), e discute como os coletores de informações BGP podem ser utilizados como fontes de informação para a identificação de sequestros de prefixo (Seção 2.3). A Seção 2.4 discute problemas de segurança do BGP e algumas abordagens recentemente propostas para saná-las. Os principais métodos de Inteligência Artificial Explicável e seus objetivos são abordados na Seção 2.5. O capítulo conclui apresentando métricas para avaliar a eficiência de modelos de IA e que permitem comparações de desempenho de diferentes modelos de aprendizado de máquina para o mesmo problema (Seção 2.6).

2.1 Sistemas Autônomos

Um Sistema Autônomo (AS, do inglês *Autonomous System*) representa uma rede, ou um conjunto de redes, administrada por uma única entidade, como uma universidade, empresa, provedor de acesso à Internet ou provedor de conteúdo. Cada AS tem autonomia para definir suas configurações, regras e políticas. Cada AS é identificado na Internet por um número único, conhecido como ASN (*Autonomous System Number*).

Os endereços IP são distribuídos aos ASes como blocos de prefixos, que incluem o endereço da rede e o número de bits usados para representar a rede. O Protocolo BGP (*Border Gateway Protocol*), detalhado na Seção 2.2, é universalmente utilizado por ASes para trocar informações de roteamento. Ele transmite informações como o prefixo de destino, o próximo salto e o *AS-path* [67]. O *AS-path* contém a sequência de ASes, representados por seus ASNs, que são percorridos até se chegar ao *AS origem* do prefixo. O AS origem é o último ASN na sequência representada pelo *AS-path*.

A gestão de números (*i.e.,* identificadores) na Internet é responsabilidade da IANA (*Internet Assigned Numbers Authority*) [3], que atribui endereços IP e números ASN aos ASes por meio dos cinco RIRs (*Regional Internet Registries*):

- **ARIN**: atende a América do Norte, (https://www.arin.net/) ;
- **LACNIC**: atende a América Latina (https://www.lacnic.net/);
- **APNIC**: atende a Ásia/Pacífico (https://www.apnic.net/);
- **AFRINIC**: atende a África (https://afrinic.net/); e
- **RIPE NCC**: atende a Europa e parte da Ásia (https://www.ripe.net/).

Essas organizações supervisionam blocos de endereços e números ASN para garantir que suas distribuições sejam justas e eficientes na Internet. Em alguns casos, os RIRs podem se dividir em Registros Nacionais de Internet, como no caso do Brasil onde a gerência dos números é realizada pelo NIC.br [29].

Os sistemas autônomos e suas interconexões formam a Internet. São milhares de ASes e de conexões que permitem que a comunicação entre um dispositivo localizado em um determinado AS acesse recursos em um outro dispositivo localizado em outro AS remoto. Cada AS divulga seus recursos, utiliza recursos de outros ASes e pode atuar como trânsito para permitir a comunicação entre ASes que não têm conexão direta. A Figura 2.1 ilustra as conexões entre ASes, destacando o uso do BGP.

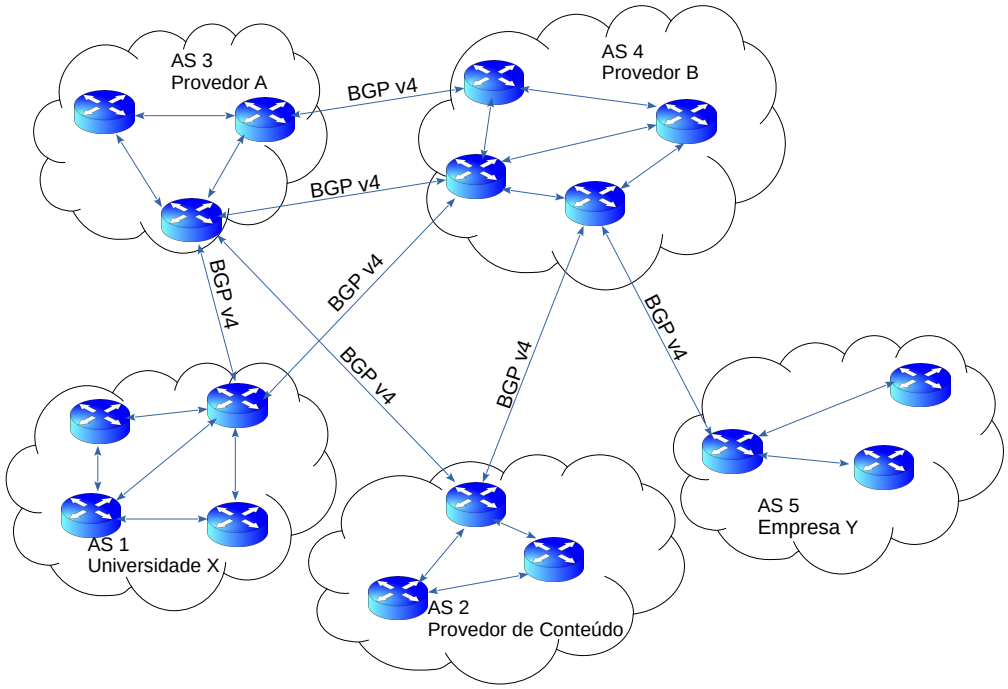


Figura 2.1: Conexão entre Sistemas Autônomos (ASes).

As trocas de informações de roteamento entre os ASes geralmente são regidas por acordos comerciais, que podem ser resumidamente classificados da seguinte maneira [21, 34]:

- ***Customer-to-Provider (c2p)* ou *Provider-to-Customer (p2c)***: Nesse tipo de acordo, o cliente (*Customer*) paga ao provedor (*Provider*) para ter trânsito através do AS do provedor, permitindo a comunicação com outros ASes. O provedor deve repassar todas as rotas conhecidas ao cliente, enquanto o cliente deve repassar ao provedor apenas as rotas para os seus próprios prefixos ou de seus clientes, caso existam. A classificação depende da direção da comunicação: do cliente para o provedor (c2p) ou do provedor para o cliente (p2c).

- ***Peer-to-Peer (p2p)***: Esse tipo de ligação ocorre entre dois ASes para permitir comunicação direta entre eles. Normalmente, eles trocam informações de rotas para acessar seus respectivos prefixos e os prefixos de seus clientes, sem que um AS tenha de pagar ao outro.

- ***Sibling-to-Sibling (s2s)***: Quando uma entidade administra dois ou mais ASes, esses ASes são chamados de irmãos (*siblings*). A ligação entre irmãos (s2s) ocorre quando esses ASes se conectam. Como a mesma entidade administra ambos, ela tem total liberdade para definir quais informações serão trocadas entre eles e pode utilizar recursos de um AS no outro, como prefixos ou comunidades BGP. Isso é comum em casos de fusão de empresas.

Os ASes podem ser classificados de acordo com suas conexões e suas principais funções. Algumas possíveis classificações são:

- ***Stub*** ou ***Single-Homed***: Um AS é considerado *stub*, ou *single-homed*, quando possui apenas uma conexão com o restante da Internet, geralmente por meio de um provedor.

- ***Multihomed***: ASes que se conectam a dois ou mais ASes, normalmente seus provedores;

- ***Internet Service Provider (ISP)***: ASes que fornecem acesso à Internet, oferecendo trânsito para que outros ASes, clientes domésticos ou empresas sem seus próprios ASes, se conectem à Internet.

- ***Transit***: ASes que fornecem trânsito, permitindo a conexão entre dois ou mais ASes que não estão diretamente conectados.

- ***Internet eXchange Point (IXP)***: Também conhecido como Ponto de Troca de Tráfego (PTT) em português, oferecem locais e infraestrutura para que outros ASes possam se conectar e trocar tráfego.

- ***Tier-1***: ASes no topo da hierarquia da Internet, que não possuem provedores e podem acessar toda a Internet sem pagar pela conexão. Normalmente, são redes de alta velocidade presentes em vários países.

Como os acordos de ligação entre os ASes não são públicos, classificá-los não é uma tarefa simples. A CAIDA (*Center for Applied Internet Data Analysis*[1]) infere e divulga

[1]https://www.caida.org/about/

mensalmente a sua classificação das conexões entre os ASes (https://publicdata.caida.org/datasets/as-relationships/), categorizando-as como p2p e c2p, além de listar os ASes considerados Tier-1. Entretanto, essa classificação não contempla todas as conexões e não possui precisão de 100%. Outros trabalhos, como o de Zhiyi Chen *et al.* [21], tentam expandir essa classificação para incluir ASes que são *siblings*.

A falta de padronização nas configurações e informações dos ASes dificulta muitos estudos sobre propriedades da Internet. Em particular, configurações de roteamento arbitrárias nos ASes podem fazer com que os roteadores não cheguem a um estado estável, levando a instabilidades de roteamento. Gao e Rexford [34] propuseram um modelo, conhecido como Gao-Rexford, que garante convergência no roteamento caso os ASes sigam algumas regras. O modelo segue a lógica de que um provedor prefere enviar o tráfego para seus clientes (ao invés de outro provedor ou *peer*), pois recebe por esse serviço. Portanto, um cliente deve divulgar a um provedor apenas os prefixos acessíveis pelo seu AS e pelos seus clientes. Se um cliente divulgar rotas de um provedor para outro, ele pode sobrecarregar sua rede com tráfego indesejado. Já um provedor deve divulgar ao cliente todas as rotas conhecidas, atuando como trânsito para o cliente acessar outros ASes, e divulgar os prefixos de seus clientes para todos os ASes conectados.

Em uma ligação *peer-to-peer* (p2p), dois ASes se conectam para melhorar o acesso mútuo e de seus clientes, sem servir de trânsito um para o outro. Nessa ligação, eles divulgam apenas seus prefixos e os de seus clientes. O tráfego p2p é preferível por ter um custo de enlace menor em comparação ao tráfego via provedor.

O modelo Gao-Rexford define uma hierarquia na Internet, com os ASes Tier-1 no topo. A divulgação de prefixos de um cliente para um provedor (c2p) sobe na hierarquia, enquanto a divulgação do provedor para o cliente (p2c) desce. Por outro lado, as ligações p2p são planas.

Ao analisar um *AS-path*, as ligações devem seguir uma sequência c2p, uma ou nenhuma ligação p2p, e uma sequência p2c, mantendo a propriedade *valley-free*, que indica ligações que levam à convergência do roteamento. A Figura 2.2 exemplifica três situações com base num *AS-path* com o caminho saindo do AS 1 até o AS 5. Em (a) há a subida na hierarquia até se chegar ao topo e depois uma descida; em (b) há uma ligação plana no topo, em ambos os casos não há a criação de um vale (*e.g.,* ligações que descem e sobem na hierarquia), logo a propriedade *valley-free* é respeitada. A Figura 2.2 (c) ilustra o caminho indo de um provedor para um cliente (ligação do AS 1 com o AS 2) e depois voltando para um provedor (ligação do AS 2 com o AS 3) criando a imagem de um vale no caminho, nesse caso há violação da propriedade.

2.2 BGP - *Border Gateway Protocol*

A versão atual do *Border Gateway Protocol* (BGP) é a versão 4, conforme definido na RFC 4271 [67], publicada em 2006, e suas atualizações (RFC 6286, RFC 6608, RFC 6793, RFC 7606, RFC 7607, RFC 7705, RFC 8212, RFC 8654 e RFC 9072). Entretanto, a primeira versão do BGP está documentada na RFC 1105 de 1989 [49]. Embora a

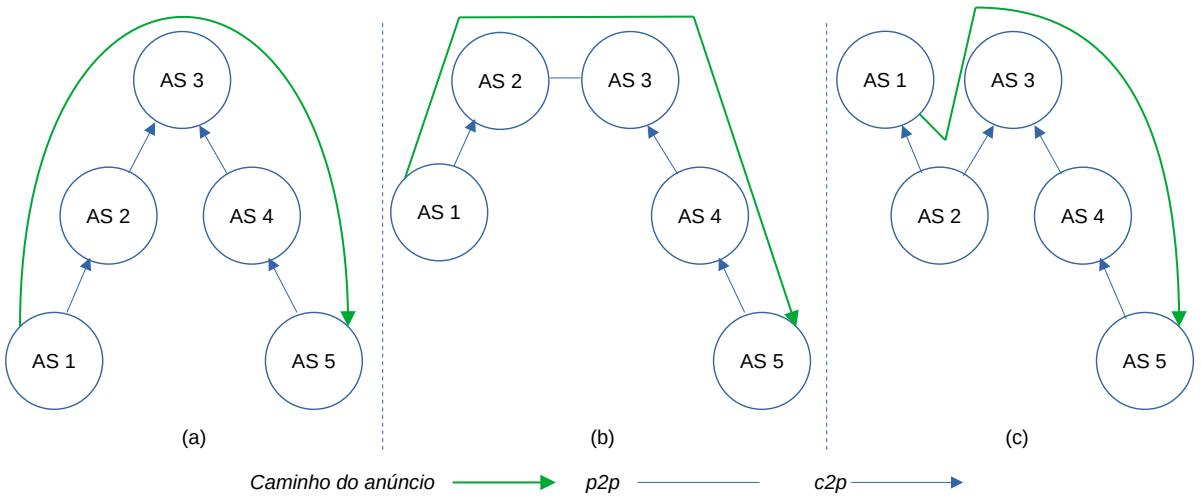


Figura 2.2: Exemplos de caminhos entre os ASes 1 e 5. As linhas direcionadas representam ligações c2p, em que a extremidade com a seta representa o provedor, e as linhas retas representam ligações p2p. Os caminhos em (a) e (b) não violam a propriedade *valley-free* enquanto o caminho em (c) viola devido a descida do provedor AS 1 para o cliente AS 2 e posterior subida do cliente AS 2 para o provedor AS 3.

implementação do protocolo e a configuração (sintaxe e comandos) variem conforme o desenvolvedor do código ou marca de equipamento, o padrão de funcionamento do BGP deve ser mantido para garantir a interoperabilidade entre equipamentos de diferentes desenvolvedores.

O protocolo BGP deve ser executado em todos os roteadores que se conectam a outro AS, podendo cada equipamento ter duas instâncias em execução. A instância `eBGP` é responsável pela conexão externa com outros ASes e está sempre em execução. A instância `iBGP` troca informações com os roteadores internos do mesmo AS e deve ser executada quando o AS possui mais de um roteador conectado a outros ASes. A Figura 2.3 ilustra essas conexões, mostrando apenas os roteadores nas fronteiras dos ASes que se conectam a outros ASes, destacando os roteadores entre os ASes X e Y. As nuvens representam as redes internas dos ASes, que podem conter outros roteadores executando diferentes protocolos de roteamento. Na ligação entre os ASes X e Y, as informações são trocadas via `eBGP` de acordo com políticas de importação e exportação de rotas. A instância `iBGP` garante a troca de rotas entre todos os roteadores internos do AS para exportação para outros ASes. A exportação de rotas segue os acordos estabelecido pelo AS com seus vizinhos.

Entre as informações trocadas via eBGP estão o prefixo da rede de destino e o *AS-path*. A Figura 2.4 exemplifica a divulgação do prefixo 192.0.2.0/24 pelo AS 1. A divulgação da rota para esse prefixo segue os passos abaixo:

1. O AS 1 divulga o prefixo 192.0.2.0/24 ao AS 2 incluindo somente seu ASN no *AS-path*. Seu ASN será sempre o mais a direita, indicando que o anúncio do prefixo partiu do seu AS;

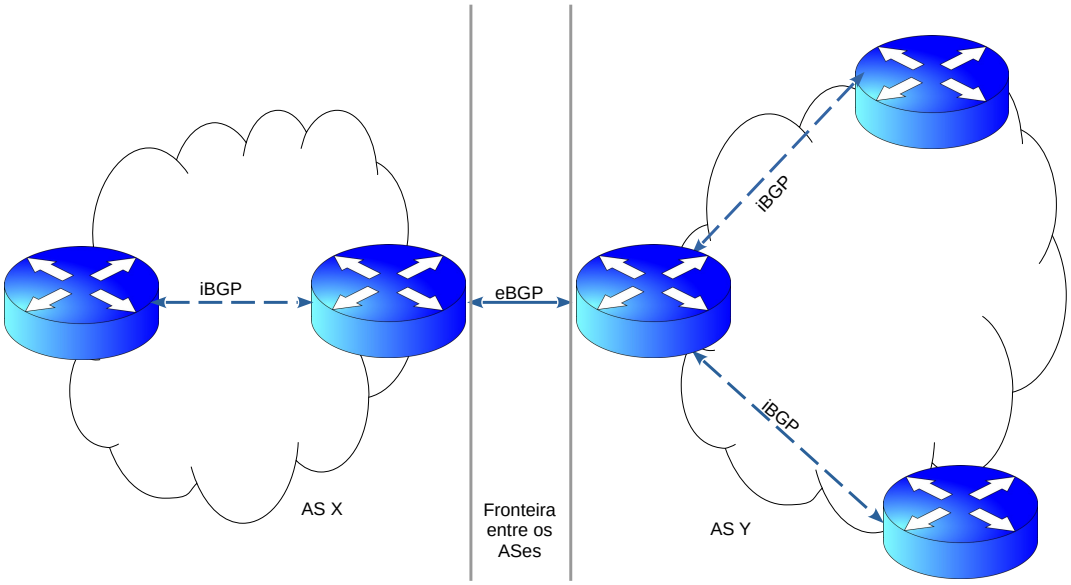


Figura 2.3: Conexão entre dois sistemas autônomos.

2. O AS 2 recebe o anúncio do AS 1 e divulga o prefixo do AS1 ao demais ASes conectados com o seguinte *AS-path* {2, 1}, com seu ASN incluído na esquerda;

3. O AS 3 recebe o anúncio do AS 2 e divulga o prefixo do AS 1 ao AS 4 com o seguinte *AS-path* {3, 2, 1};

4. O AS 4 recebe os anúncios relativos ao prefixo do AS 1 oriundo do AS 2 e do AS 3. O anúncio selecionado nesse exemplo foi o recebido do AS 2 por ter um *AS-path* menor, um caminho mais curto, então o AS 4 faz o anúncio do prefixo ao AS 5 com o seguinte *AS-path* {4, 2, 1}. Ele também fará o anúncio ao AS 3, mas como o *AS-path* recebido pelo AS 3 do AS 2 é menor, ele não altera sua tabela de roteamento.

5. O AS 5 receberá o anúncio que indica que para chegar a um endereço da rede 192.0.2.0/24 ele deve enviar pacotes para o AS 4 que repassará ao AS 2 para depois chegar ao AS 1, responsável pelo anúncio.

   Obs. Para facilitar o entendimento, este exemplo não considera políticas sofisticadas de importação e exportação de rotas.

Além de auxiliar na escolha de caminhos, os ASNs no *AS-path* são usados para evitar *loops* na tabela de roteamento. Quando um roteador recebe um anúncio que contém seu próprio ASN no *AS-path*, ele descarta essa rota para evitar um *loop*. Segundo a RFC 4271 [67], o algoritmo de seleção da melhor rota para um prefixo do BGP deve seguir a ordem abaixo, avançando para o próximo critério apenas em caso de empate:

1. ***Local preference***: O administrador do AS pode atribuir um valor de *local preference* às rotas recebidas com base nas políticas internas. A rota com o valor de *local preference* mais alto será escolhida.

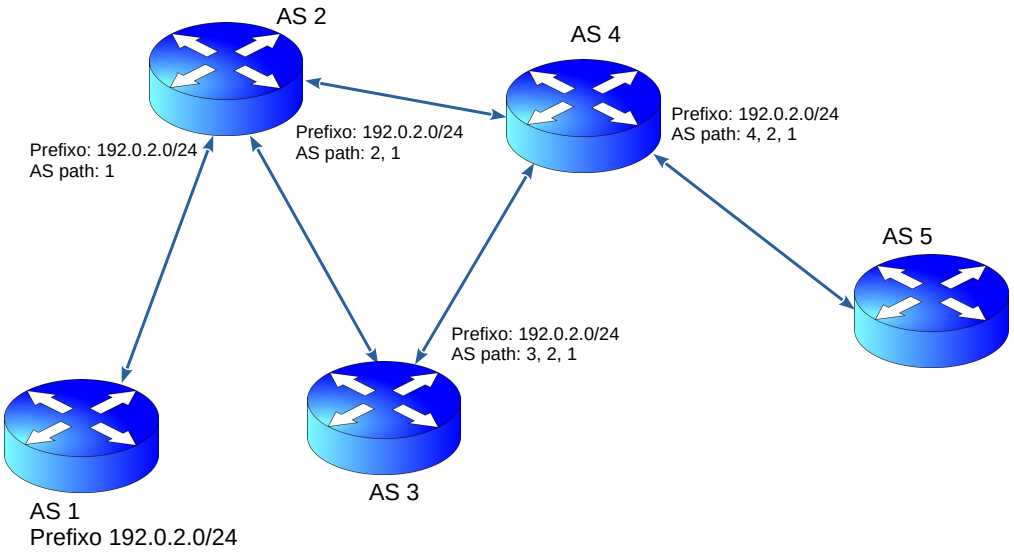


Figura 2.4: Exemplo de divulgação de um prefixo pelo AS 1 até todos os ASes receberem o anúncio.

2. **Comprimento do *AS-path***: A rota com o menor número de ASNs no *AS-path* será selecionada.

3. **Número de origem**: A rota com o menor número de origem será preferida. Este número pode ser 0 (rota aprendida internamente via IGP), 1 (rota aprendida externamente via EGP), ou 2 (rota com informações incompletas ou aprendida de outra forma).

4. ***Multiple Exit Discriminator* (MED)**: Quando um AS possui várias conexões com outro AS, o MED pode ser configurado para influenciar a escolha da rota. A rota com o menor valor de MED será escolhida.

5. **Entre iBGP e eBGP**: Rotas recebidas via eBGP têm prioridade sobre rotas recebidas via iBGP.

6. **Custo para o próximo salto**: A rota com o menor custo para alcançar o próximo salto será preferida.

7. **Identificador BGP**: A rota com o maior identificador BGP será escolhida. O identificador BGP é configurado como um endereço IPv4, geralmente o endereço de uma interface de *loopback*, e é o mesmo para todas as interfaces do roteador.

8. **Endereço do *peer***: Em caso de empate, a rota com o menor endereço de conexão BGP do outro AS será selecionada.

Políticas de importação permitem priorizar ou rejeitar rotas com base no AS de conexão ou em um ASN presente no *AS-path*, por exemplo. Os critérios mencionadas

acima são seguidos somente após a rota ter sido aceita pela política de importação do AS. A rota escolhida pelo algoritmo de seleção da melhor rota do BGP é adicionada à tabela de roteamento do equipamento e usada no plano de dados para decidir o encaminhamento dos pacotes, indicando para qual equipamento o pacote deve ser enviado.

Os critérios de desempate para escolha de rotas apresentados aqui estão de acordo com a RFC 4271 [67], que define o funcionamento do protocolo. No entanto, a codificação e implementação do protocolo no equipamento dependem do desenvolvedor, e alguns fabricantes fazem pequenas alterações que afetam apenas seus próprios dispositivos. Por exemplo, a Cisco possui 13 critérios de desempate para rotas, incluindo algumas além das existentes na RFC, como um peso atribuído à rota e a preferência pelo anúncio recebido primeiro [24].

Em uma tabela de roteamento, pode haver mais de um prefixo que corresponda ao endereço de destino. Por exemplo, se o destino a ser alcançado é o endereço IPv4 `10.10.10.2`, e a tabela de roteamento contém os prefixos `10.10.10.0/24` e `10.10.0.0/16`, ambos abrangem o endereço de destino. Nesse caso, a rota através do prefixo `10.10.10.0/24` será escolhida por ser mais específica. A escolha do prefixo mais longo é definida pelo protocolo IP (*Internet Protocol*) [5].

```
rviews@rviews-vm> show route 8.8.8.8

inet.0: 948361 destinations, 1762577 routes (948353 active, 0 holddown, 15 hidden)
+ = Active Route, - = Last Active, * = Both

8.8.8.0/24          *[BGP/170] 25w3d 12:30:28, localpref 100, from 164.113.199.100
                      AS path: 1299 15169 I
                     > to 164.113.193.222 via ge-0/0/0.0
                     [BGP/170] 21:04:25, MED 951, localpref 100, from 164.113.199.110
                      AS path: 11164 15169 I
                     > to 164.113.193.222 via ge-0/0/0.0

small.inet.0: 167347 destinations, 219225 routes (167347 active, 0 holddown, 0 hidden)
+ = Active Route, - = Last Active, * = Both

8.0.0.0/12          *[BGP/170] 3d 13:29:20, MED 11020, localpref 100, from 164.113.199.105
                      AS path: 174 3356 I
                     > to 164.113.193.222 via ge-0/0/0.0
                     [BGP/170] 5w0d 06:01:29, localpref 100, from 164.113.199.110
                      AS path: 6939 3356 I
                     > to 164.113.193.222 via ge-0/0/0.0
```

Figura 2.5: Visualização do *AS-path* em um servidor de rotas com acesso público para consulta.

A Figura 2.5 mostra um exemplo de rotas para o endereço IPv4 `8.8.8.8` obtidas de um servidor de rotas público. Nela, dois prefixos atendem ao endereço: `8.8.8.0/24` e `8.0.0.0/12`. Para ambos os prefixos, há dois caminhos possíveis. Os pacotes destinados ao prefixo `8.8.8.0/24` podem seguir pelo *AS-path* {1299, 15169} ou pelo *AS-path* {11164, 15169}, enquanto os pacotes destinados ao prefixo `8.0.0.0/12` podem seguir pelo *AS-path* {174, 3356} ou pelo *AS-path* {6939, 3356}. A seleção da rota ativa para cada prefixo seguirá os critérios mencionados anteriormente. A figura também mostra a origem de

cada prefixo: o prefixo `8.8.8.0/24` é originado pelo ASN 15169, e o prefixo `8.0.0.0/12` pelo ASN 3356. Nesse caso, é provável que o prefixo `8.0.0.0/12` seja de propriedade do ASN 3356, que cedeu o prefixo `8.8.8.0/24` para o ASN 15169. No entanto, isso também poderia indicar o sequestro de um prefixo mais específico pelo ASN 15169. Os tipos de sequestros são discutidos no Capítulo 3.

2.3 Coletores de Informações BGP

Os projetos Route Views [60] e RIPE RIS [66] coletam atualizações do BGP e as RIBs (*Routing Information Base*) de roteadores em vários pontos da Internet e disponibilizam essas informações de forma livre, incluindo dados históricos e quase em tempo real. Ambos os projetos funcionam de maneira semelhante: coletores distribuídos pela Internet recebem informações do BGP de vários ASes, que são armazenadas e acessíveis através dos sites dos projetos. Esses dois coletores são os mais utilizados nos estudos e também são usados neste trabalho.

As informações coletadas pelos projetos Route Views e RIPE RIS são disponibilizadas de duas formas: arquivos RIB, que são "fotos" das rotas conhecidas pelo roteador no momento da coleta, e arquivos de *updates*, que contêm atualizações das rotas anunciadas (inclusão, alteração ou exclusão). Atualmente, os arquivos RIB são gerados a cada duas horas no Route Views e a cada oito horas no RIPE RIS. Os arquivos de *updates* são gerados a cada quinze minutos no Route Views e a cada cinco minutos no RIPE RIS. Esses arquivos são compactados e requerem aplicativos para análise, como `bgpdump`, `bgpreader`, `bgpscanner` e `bgpkit`. Algumas dessas ferramentas estão disponíveis como bibliotecas para linguagens de programação, como `pybgpkit` para Python. O presente trabalho utilizou o `bgpscanner` para análise dos arquivos baixados, devido à sua velocidade na extração de informações. O `bgpreader` foi utilizado para análises pontuais, pois permite a aplicação de filtros que geram apenas as informações selecionadas.

O uso de coletores para analisar o comportamento dos anúncios BGP é frequente entre pesquisadores. O site do Route Views, por exemplo, lista mais de mil citações (https://www.routeviews.org/routeviews/index.php/papers/). Pesquisas como a de Milolidakis *et al.* [56] analisam a capacidade dos coletores de detectar comportamentos maliciosos de ASes na Internet. Eles mostram que é possível ocultar ataques BGP dos coletores, mas isso requer conhecimento detalhado sobre a localização dos pontos de coleta e as políticas de roteamento, além de limitar o alcance do ataque. Outro estudo [72] mostra que ataques que atingem menos de 1% da Internet podem não ser detectados, mas a maioria das atividades maliciosas não observadas teve um alcance médio de apenas 0,2%, destacando a eficácia dos coletores públicos na detecção de atividades maliciosas que utilizem BGP.

2.4 Segurança de BGP

BGP não possui mecanismos de segurança e autenticação para impedir que um AS use um ASN ou prefixo de rede que não lhe pertence. O protocolo baseia-se na confiança de que

cada AS divulgará apenas seus próprios prefixos [23]. Isso permite que qualquer entidade na Internet anuncie qualquer prefixo de rede sem restrições. Alterações nos anúncios recebidos e repassados, como mudanças no *AS-path*, também são possíveis. Algumas alterações são intencionais, como o *prepend*, em que um AS adiciona várias vezes seu próprio ASN no *AS-path* para tornar o caminho menos preferível. No entanto, ASNs não pertencentes ao caminho real também podem ser adicionados. A falta de segurança em BGP permite ataques como o sequestro de prefixo (*Prefix Hijacking* ou *BGP hijacking*), em que um AS anuncia um prefixo que não lhe pertence.

A falta de segurança em BGP tem sido abordada por diversos estudos que propõem soluções para melhorar a segurança do protocolo. A primeira e mais eficaz medida é a implementação de políticas de importação e exportação de rotas por todos os ASes. Lychev *et al.* [50] destacam que essas políticas permitem que um AS aceite apenas as rotas de seus clientes, prevenindo a propagação de prefixos indevidamente anunciados. Pode-se afirmar que essa seria a medida de segurança mais simples e eficaz contra divulgação incorreta de prefixos na Internet. Entretanto, essa prática não é universalmente adotada, como evidenciado por incidentes contínuos de sequestro de prefixos. Por exemplo, em 4 de abril de 2010, um AS da China Telecom sequestrou mais de 50.000 prefixos, afetando 15% da Internet na época de mais de 70 países diferentes [84]. Em 2021, o relatório de segurança BGP do MANRS (*Mutually Agreed Norms for Routing Security*) registrou sequestros afetando até 30.000 prefixos e vazamentos de rotas atingindo 167.000 prefixos [77], confirmando que essas políticas ainda não são amplamente aplicadas.

Uma das melhores propostas para melhorar a segurança do BGP é o uso de RPKI (*Resource Public Key Infrastructure*), que utiliza uma infraestrutura de chaves públicas para validar prefixos e seus ASNs de origem. A definição atualizada da RPKI está na RFC 8210 [11]. Sua implementação envolve duas partes:

- ROA (*Route Origin Authorization*): Após obter um certificado do RIR responsável, um AS pode definir quais ASNs estão autorizados a anunciar seus prefixos. Essa autorização é assinada digitalmente para garantir sua validade.

- ROV (*Route Origin Validation*): Esta etapa valida se um prefixo recebido por um roteador é proveniente de um ASN autorizado a divulgá-lo.

Se RPKI fosse amplamente adotado, a divulgação de prefixos por ASNs não autorizados teria um impacto reduzido. No entanto, a implementação de RPKI ainda é baixa. Sermpezis *et al.* [71] mostram que 71% dos operadores de rede não implementaram RPKI, citando custos e aumento da complexidade como principais razões. De acordo com o site RPKI Monitor [59], em 13 de fevereiro de 2024, 47,82% dos endereços IPv4 e 50,50% dos endereços IPv6 estavam cobertos pelo ROA, indicando progresso, mas ainda há muito a ser feito.

Uma nova técnica de segurança para o BGP, chamada ASPA (*Autonomous System Provider Authorization* — Autorização de Provedor de Sistema Autônomo), está sendo discutida para aproveitar a estrutura de RPKI na validação das conexões entre ASes. Embora seu funcionamento ainda esteja em estudo e sua adoção seja limitada, a implementação de ASPA permitirá aos ASes verificar a validade das conexões nos

AS-paths, descartando rotas com *AS-paths* forjados ou resultantes de vazamentos de rotas [4].

O BGPsec, definido na RFC 8205 [46], é outra iniciativa para aumentar a segurança do BGP, permitindo a validação do *AS-path* através de assinaturas digitais. O principal objetivo do BGPsec é permitir que um AS valide um *AS-path* recebido, assegurando que ele não foi alterado. O seu funcionamento consiste na assinatura digital do *AS-path* por todos os ASes no caminho. Assim o próximo AS a receber o anúncio consegue validar o caminho anterior a ele. Entre os motivos de sua não implementação, podem-se destacar a necessidade de uma maior capacidade de processamento nos roteadores para o uso de criptografia, necessária para assinatura e validação das rotas, e a exigência de adesão universal.

Outro protocolo desenvolvido para aumentar a segurança do BGP é o ROV++ [58], que aprimora o ROV existente na RPKI e pode ser implementado em três versões:

- Versão 1: Além de selecionar apenas rotas confiáveis e validadas, descarta pacotes destinados a rotas não seguras.
- Versão 2: Divulga rotas não seguras aos seus clientes, redirecionando o tráfego para descartá-lo.
- Versão 3: Protege uma área maior ao observar um sequestro, divulgando a rota insegura para todos e descartando os pacotes.

Segundo Morillo *et al.* [58], o descarte de pacotes destinados a rotas inseguras é necessário para eliminar sequestros ocultos. Esse tipo de sequestro ocorre quando há pelo menos dois anúncios para o mesmo endereço de destino, um legítimo e outro não, ambos com o mesmo próximo AS. Isso pode acontecer em casos de sequestro de um prefixo mais específico, em que o pacote ainda pode ser direcionado ao AS sequestrador, mesmo descartando-se o anúncio ilegítimo.

Apesar das diversas soluções propostas para resolver a falta de segurança no BGP, nenhuma atingiu uma implementação suficientemente ampla para solucionar os problemas de segurança. Isso é evidenciado pelos recorrentes sequestros e vazamentos de rotas. As justificativas para a falta de uma solução amplamente implementada incluem os custos de capacitação dos operadores de rede e a substituição de equipamentos devido à necessidade de maior processamento para uso de criptografia. Além disso, alguns ASes envolvidos em atividades maliciosas não desejam que essas vulnerabilidades sejam corrigidas. Konte *et al.* [42] demonstram a existência de ASes conhecidos como *Bulletproof Hosting ASes*, cujo objetivo é hospedar sistemas maliciosos e garantir sua disponibilidade.

2.5 Inteligência Artificial Explicável

Nos últimos anos, aprendizado de máquina tem sido utilizado para resolver diversos problemas, empregando modelos cada vez mais complexos baseados em redes neurais profundas, florestas aleatórias e LLMs (*Large-Language Models*). Esses modelos são considerados caixas-pretas (*black boxes*) e geralmente são preferidos aos modelos

interpretáveis, como árvores de decisão, devido ao seu desempenho superior. Entretanto, apesar da atenção que recebem, operadores de rede ainda hesitam em adotar modelos caixas-pretas em situações críticas, como bloqueio de tráfego, por não entenderem como as decisões são tomadas [40, 83]. Além disso, esses modelos complexos exigem uma quantidade maior de *features* que podem depender de informações de fontes nem sempre confiáveis e demandam maior tempo de processamento para computá-las.

Diversos esforços na área de Inteligência Artificial Explicável (XAI) propõem métodos para explicar ou examinar a validade das classificações geradas por modelos caixa-preta e determinar a relevância das *features* no processo de decisão. As abordagens de XAI podem ser divididas em dois grupos principais:

- **Explicabilidade Local**: a explicabilidade local são técnicas que fazem a análise de uma inferência buscando identificar qual o peso, ou seja, a importância de cada *feature* na inferência do modelo em cada caso. Elas buscam iluminar as decisões individuais realizadas. O `Lime` [68] é uma das ferramentas que pode ser utilizada para se obter a explicabilidade local;

- **Explicabilidade Global**: a explicabilidade global busca descrever como o modelo toma suas decisões de forma global, não se limitando a casos individuais. As técnicas utilizadas buscam a criação de modelos humanamente compreensíveis baseados no modelo caixa-preta, como na criação de uma DT (*Decision Tree* - Árvore de Decisão) ou de regras de decisão. Esses modelos interpretáveis podem ser compreendido por um especialista no domínio do problema [40, 44].

A ferramenta Trustee [40], destinada à explicabilidade global, gera árvores de decisão de alta fidelidade para explicar o funcionamento de um modelo caixa-preta. Cada árvore de decisão é gerada utilizando aprendizado por imitação (*imitation learning*) [44] usando uma dinâmica professor-aluno em que um conjunto de amostras é passado para o modelo caixa-preta classificar. As amostras classificadas pelo modelo caixa-preta (professor) são usadas para treinar a árvore de decisão (aluno). O objetivo é gerar árvores de decisão que imitam o funcionamento do modelo caixa-preta. Assim, um especialista pode analisar as árvores de decisão para verificar se as decisões tomadas pelo modelo estão de acordo com o conhecimento existente no domínio do problema. Portanto, a árvore de decisão gerada por Trustee atua como um modelo substituto compreensível para o especialista. Trustee já foi utilizada para identificar problemas em modelos caixas-pretas em diferentes domínios de aplicação, especialmente na área de segurança de redes [6, 40].

2.6 Métricas de Aprendizado de Máquina

Diversas métricas podem ser usadas para avaliar o desempenho de um modelo de aprendizado de máquina, como, por exemplo, precisão e *recall*. Como as classificações dos modelos analisados são binárias (*e.g.,* legítimos vs. suspeitos ou negativos vs. positivos), o cálculo das métricas exige o conhecimento dos seguintes valores:

- **Verdadeiros Positivos (VP)**: Quantidade de ASes ou enlaces suspeitos corretamente identificados.

- **Verdadeiros Negativos (VN)**: Quantidade de ASes ou enlaces legítimos corretamente identificados.
- **Falsos Positivos (FP)**: Quantidade de ASes ou enlaces legítimos incorretamente classificados como suspeitos.
- **Falsos Negativos (FN)**: Quantidade de ASes ou enlaces suspeitos incorretamente classificados como legítimos.

Com base nesses valores, as seguintes métricas podem ser calculadas:

- **Taxa de Verdadeiros Positivos (TVP)**: Representa a proporção de ASes ou enlaces suspeitos identificados corretamente, calculada como $\text{TVP} = \frac{\text{VP}}{\text{TP}}$, em que TP é o total de positivos. Valores próximos de um indicam melhor desempenho.
- **Taxa de Falsos Positivos (TFP)**: Indica a proporção de ASes ou enlaces legítimos incorretamente classificados como suspeitos, calculada como $\text{TFP} = \frac{\text{FP}}{\text{TN}}$, em que TN é o total de negativos. Valores menores são desejáveis.
- **Taxa de Verdadeiros Negativos (TVN)**: Mede a proporção de ASes ou enlaces legítimos corretamente classificados, calculada como $\text{TVN} = \frac{\text{VN}}{\text{TN}}$. Valores próximos de um indicam maior taxa de acerto.
- **Taxa de Falsos Negativos (TFN)**: Representa a proporção de ASes ou enlaces suspeitos classificados incorretamente como legítimos, calculada como $\text{TFN} = \frac{\text{FN}}{\text{TP}}$. Valores menores são preferíveis.
- **Precisão**: Avalia a proporção de previsões corretas dentro de cada classe. Para a classe positiva (suspeitos), é calculada como

$$\text{Precisão}_{\text{P}} = \frac{\text{VP}}{\text{VP+FP}}$$

  Para a classe negativa (legítimos), é calculada como

$$\text{Precisão}_{\text{N}} = \frac{\text{VN}}{\text{VN+FN}}$$

- ***Recall***: Mede a proporção de elementos corretamente identificados em relação ao total de elementos na classe. Para a classe positiva (suspeitos), é equivalente à métrica TVP

$$\text{Recall}_{\text{P}} = \frac{\text{VP}}{\text{TP}}$$

  Para a classe negativa (legítimos), equivale à métrica TVN

$$\text{Recall}_{\text{N}} = \frac{\text{VN}}{\text{TN}}$$

- ***F1-Score***: Calcula a média harmônica entre as métricas de precisão e *recall* para cada classe, representando o equilíbrio entre elas. É dado por:

$$\text{F1-Score} = 2 \cdot \frac{\text{Precisão} \cdot \text{Recall}}{\text{Precisão} + \text{Recall}}$$

- **Acurácia**: Indica a proporção total de classificações corretas, mas pode ser pouco representativa em conjuntos desbalanceados. É calculada como:

$$\text{Acurácia} = \frac{\text{VP}+\text{VN}}{\text{TP}+\text{TN}}$$

- ***Matthews Correlation Coefficient (MCC)***: Avalia a eficiência do modelo em classes balanceadas ou desbalanceadas. Seu valor varia de -1 a 1, sendo que valores próximos de 1 indicam alta eficiência. Sua fórmula é:

$$\text{MCC} = \frac{\text{VP} \cdot \text{VN} - \text{FP} \cdot \text{FN}}{\sqrt{(\text{VP}+\text{FP}) \cdot (\text{VP}+\text{FN}) \cdot (\text{VN}+\text{FP}) \cdot (\text{VN}+\text{FN})}}$$

Capítulo 3

Sequestro de Prefixo

Como o BGP não possui mecanismos de segurança e validação de anúncios recebidos, um AS pode divulgar prefixos que não lhe pertencem ou fazer alterações em um *AS-path* tanto de forma maliciosa ou por erros de configuração. Um *sequestro de prefixo* ocorre quando um AS divulga um prefixo que não lhe pertence sem a autorização do AS proprietário do prefixo. Este capítulo apresenta as principais causas para sequestros de prefixo, os tipos conhecidos, suas classificações e seus prováveis alcances.

3.1 Principais Causas

As motivações para sequestros maliciosos de prefixo na Internet são das mais variadas e, normalmente, visam encobrir operações ilegítimas, se passar pela rede de origem ou evitar que tráfego legítimo chegue ao destino correto do prefixo sequestrado. Uma ocorrência frequente de sequestro de prefixo é para a utilização dos endereços IP do prefixo para envio de *spam*, pois endereços utilizados para esse propósito costumam ser bloqueados nos servidores de e-mail e compartilhados em listas de bloqueio [23,43,50,58,74,81,84]. Assim, os grandes geradores de *spam* não utilizam seus próprios endereços para não entrarem em listas de bloqueio e perderem suas capacidades de envio de e-mail.

Há também relatos de sequestros de prefixo para obtenção de certificados digitais em nome do proprietário do prefixo [35], disponibilização de páginas falsas para roubo de dados [85] e roubo de informações dos usuários do AS vítima [23, 73]. Esse tipo de sequestro tem potencial de causar prejuízos financeiros significativos em suas vítimas. Por exemplo, um sequestro em 2 de fevereiro de 2022 resultou em roubo de cerca de 1,9 milhão de dólares em criptomoedas da Klayswap [78] enquanto outro em 17 de agosto de 2022 fez com que usuários da plataforma Celer Bridge perdessem cerca de 235 mil dólares [41]. Por fim, alguns sequestros desviam o tráfego para um AS malicioso com o único propósito de descartar os pacotes (*i.e., blackhole*) destinados ao AS de origem e tornar indisponíveis os seus serviços [75].

Um sequestro de prefixo nem sempre é malicioso. Há vários provedores de serviço na Internet que oferecem serviços de mitigação de ataques DDoS (*Distributed Denial-of-Service Attacks*) por meio de sequestro de prefixo. Durante um ataque DDoS, diversos dispositivos, normalmente pertencentes a uma *botnet* [25], enviam dados

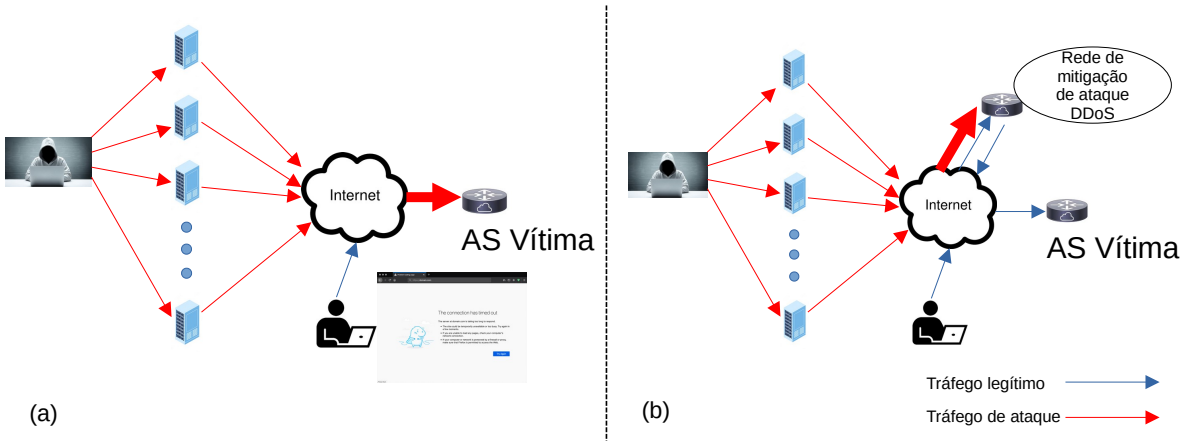


Figura 3.1: Em (a) um ataque de DDoS sem um provedor de serviço de proteção e em (b) um ataque com a atuação de um provedor de serviço de proteção.

simultaneamente para o AS vítima com a intenção de sobrecarregar os seus recursos e tornar seus serviços indisponíveis. Ao detectar o ataque, o provedor de serviços passa a divulgar o prefixo que está sob ataque como se fosse seu para evitar que o tráfego chegue até a rede da vítima. Normalmente, esse tipo de provedor possui várias conexões com outros ASes e consegue redirecionar rapidamente todo, ou uma fração significativa, do tráfego para sua infraestrutura de mitigação que é capaz de descartar pacotes maliciosos e redirecionar o tráfego legítimo para o AS vítima por um enlace dedicado ou por meio de tunelamento, por exemplo. A Figura 3.1(a) ilustra uma rede sem proteção de ataques DDoS enquanto a Figura 3.1(b) mostra a atuação do provedor contratado para a mitigação de ataques DDoS.

Um sequestro também pode ser causado por erros de configuração, como, por exemplo, a digitação incorreta de um prefixo. Apesar de não ser intencional, ele também causa grandes transtornos para o AS proprietário do prefixo, pois o tráfego para o prefixo pode ser redirecionado e não chegar até o destino correto. Um exemplo ocorreu em maio de 2016 em que o AS 203959 passou a anunciar o prefixo `191.86.129.0/24` ao invés do seu prefixo `191.96.129.0/24`, o que possivelmente indica um erro de digitação com a troca do dígito 9 pelo 8 no segundo octeto. Esse sequestro foi descoberto pela origem do prefixo anunciado [23] e rapidamente corrigido. Normalmente, os sequestros causados por erros de configuração fazem com que o tráfego não chegue ao AS destino, pois os pacotes são descartados pelo AS que cometeu o erro.

3.2 Classificação

Como um sequestro de prefixo pode ocorrer de diferentes maneiras, alguns trabalhos propõem classificações para os tipos conhecidos de sequestro. As classificações se dividem com relação a mudanças no *AS-path* ou ao prefixo sequestrado.

Quando um prefixo sequestrado é anunciado com o AS sequestrador como origem e

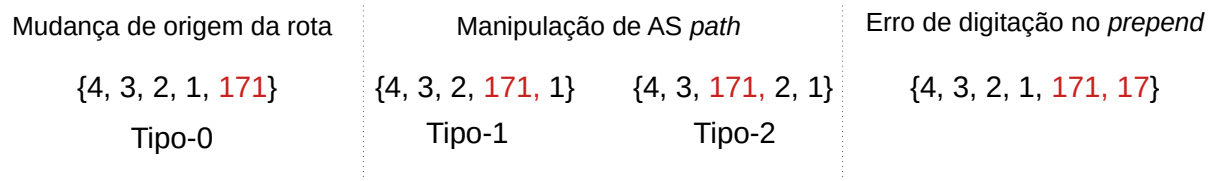


Figura 3.2: Exemplos de classificação de diferentes tipos de sequestro manifestados nos *AS-paths* em que o AS 171 é o sequestrador. Os sequestros de Tipo-`X` mostram o AS 171 em diferentes posições (`X`=0, 1 ou 2) no *AS-path*. O quarto exemplo ilustra um erro de digitação quando o objetivo era duplicar o AS 171 no *AS-path*, mas acidentalmente o ASN adicionado foi o 17 ao invés do 171.

também pelo seu proprietário legítimo, ocorre o fenômeno conhecido como MOAS (*Multiple Origin AS*), em que múltiplos ASes são identificados como origem para o mesmo prefixo. Cho *et al.* [23] clasificam esse tipo de sequestro como mudança de origem da rota (*route origin change*), enquanto Sermpezis *et al.* [72] o classificam como Tipo-0, indicando que não há nenhum ASN anterior ao ASN do sequestrador no *AS-path*.

Outro tipo de sequestro ocorre quando o sequestrador forja ou manipula o *AS-path*. Cho *et al.* [23] classificam esse tipo de sequestro como de manipulação de *AS-path* apenas. Por outro lado, Sermpezis *et al.* [72] o classificam como de Tipo-`X`, em que `X` corresponde à quantidade de ASNs introduzidos ou manipulados antes do ASN do sequestrador no *AS-path*. Erros de digitação tanto no prefixo quanto em configurações de política de roteamento também podem ocasionar sequestros de prefixo ou modificar o *AS-path*, fazendo com que ele pareça forjado. A Figura 3.2 ilustra alguns tipos de sequestro de prefixo e as consequentes mudanças nos *AS-paths*. Os três primeiros exemplos ilustram sequestros de Tipo-`X`, para diferentes valores de `X`, e o quarto ilustra um erro de digitação na configuração de *prepend*. Neste caso, o objetivo era duplicar o ASN 171 para inflar artificialmente o *AS-path* e diminuir a prioridade da rota, mas um erro de digitação introduziu o ASN 17 ao invés do 171.

Com relação ao prefixo sequestrado, um sequestro pode ocorrer com diferentes cenários. Primeiro, ele pode ser um sequestro de prefixo que corresponde exatamente ao prefixo anunciado pelo AS proprietário. Nesse caso, o AS sequestrador anuncia um prefixo de mesmo comprimento que o AS proprietário. Segundo, o prefixo sequestrado pode ser mais específico, ou seja, representar uma subrede do prefixo original. Por fim, o prefixo sequestrado pode não estar alocado ou o AS proprietário não o anuncia regularmente, caracterizando uma técnica de sequestro conhecida como *IP prefix squatting* [72].

3.3 Alcance

Um sequestro de prefixo pode impactar no tráfego de um número variável de ASes, desde alguns poucos até todos os ASes da Internet, dependendo de como o sequestro é realizado.

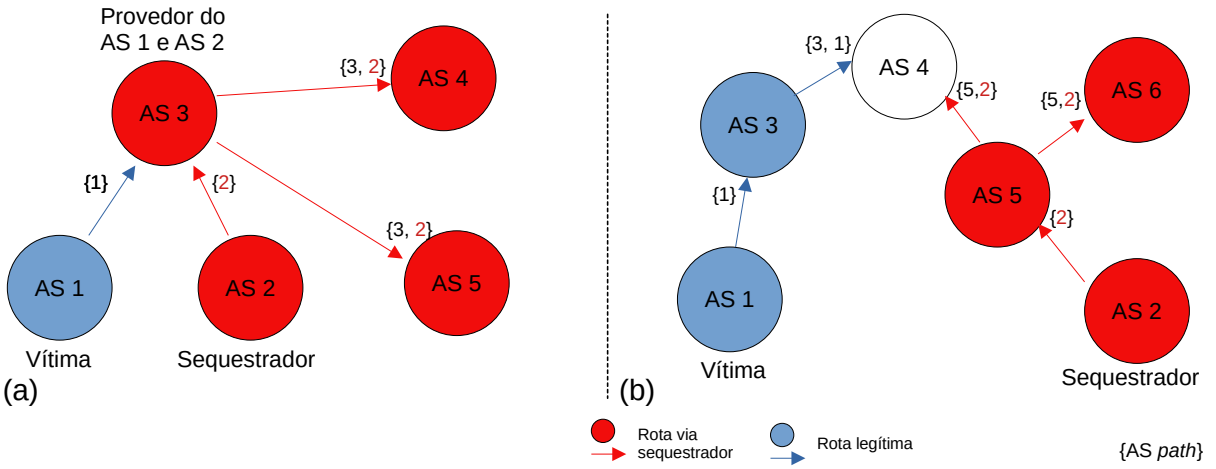


Figura 3.3: Exemplo do impacto do sequestro com base na localização dos ASes vítima e sequestrador.

Por exemplo, se o prefixo sequestrado não estiver sendo anunciado por nenhum outro AS, o sequestro pode se propagar por toda a Internet, independente da forma do sequestro. Isso ocorre porque não há mais de uma origem anunciando o prefixo e, portanto, a rota escolhida poderá levar sempre à mesma origem, ou seja, ao AS sequestrador. Uma exceção é quando o AS sequestrador inclui outros ASNs no *AS-path* para gerar um sequestro do Tipo-`X`. Neste caso, todos os ASes que tiveram seus ASNs incluídos pelo sequestrador rejeitarão a rota anunciada para evitar *loops* de roteamento utilizando o mecanismo padrão de prevenção de *loops* do protocolo BGP.

Quando o sequestrador anuncia um prefixo idêntico ao do AS proprietário (*i.e.,* de mesmo comprimento), o alcance do sequestro varia muito em função da conectividade do AS sequestrador e também da conectividade do AS proprietário, pois anúncios das duas origens vão participar do processo de escolha de rotas do protocolo BGP em cada AS. Como cada AS repassa somente a rota escolhida, alguns ASes receberão somente rotas para o AS proprietário enquanto outros receberão somente rotas para o AS sequestrador. Por outro lado, quando o sequestrador anuncia um prefixo mais específico do que aquele anunciado pelo AS proprietário, o alcance do sequestro é global, pois cada AS na Internet terá uma rota para o prefixo original e outra para o prefixo mais específico. Como o protocolo IP seleciona a rota com o prefixo mais longo, todo o tráfego destinado para o prefixo sequestrado será desviado para o AS sequestrador.

O alcance de um sequestro do Tipo-`X` depende diretamente do valor de `X`, ou seja, da quantidade de ASNs incluídos no *AS-path*. Como um dos primeiros critérios de seleção de rotas de BGP é o comprimento do *AS-path*, valores maiores de `X` reduzem significativamente o alcance do sequestro, pois tornam o *AS-path* maior e, consequentemente, diminuem a prioridade da rota anunciada. Segundo Sermpezis *et al.* [72], os sequestros com *AS-path* forjados com mais de 2 ASNs possuem um impacto muito limitado, sendo que um sequestro de Tipo-3 impacta na média em torno de 4% da Internet enquanto um de Tipo-4 somente 1%.

Dois cenários extremos ajudam a ilustrar como as localizações dos ASes sequestrador

e vítima em relação a outros ASes influenciam no alcance de um sequestro:

- **Primeiro cenário**: os ASes vítima e sequestrador possuem somente uma conexão com a Internet e com o mesmo provedor. Caso o provedor escolha a rota para o prefixo sequestrado via AS do sequestrador, ele irá divulgar a rota com o sequestro para toda a Internet e a rota legítima não será vista por nenhum AS. A Figura 3.3(a) ilustra essa situação. Por outro lado, se o provedor escolher a rota legítima, o sequestro não terá impacto algum.

- **Segundo cenário**: o AS vítima está distante do AS sequestrador. Caso o sequestro seja sem manipulação do *AS-path*, aqueles ASes que estão mais próximos da vítima não serão afetados e somente os mais próximos do sequestrador serão impactados pelo sequestro. A Figura 3.3(b) ilustra essa situação. No exemplo da figura, somente o AS 4 receberá dois anúncios com *AS-path* de tamanho igual e terá que selecionar um dos dois com base nos critérios do protocolo BGP que foram discutidos no Capítulo 2.

Os exemplos de alcance de um sequestro discutidos acima não consideram políticas de roteamento que os ASes podem implementar (*e.g.,* valores diferentes de LocalPref ou filtros de importação e exportação de rotas) e nem o uso de mecanismos de segurança propostos para o protocolo BGP. Por exemplo, sequestros de Tipo-`0` podem ser evitados com o uso de RPKI [11]. Por outro lado, os sequestros ilustrados na Figura 3.3 podem ser evitados com o uso de políticas de importação de rotas do protocolo BGP e o uso de filtros que indicam quais prefixos um AS (*e.g.,* provedor) pode aceitar de seus vizinhos (*e.g.,* clientes). No exemplo da Figura 3.3(a), o sequestro ilustrado não terá efeito algum caso o AS 3 utilize uma política de importação que aceite somente os prefixos conhecidos de seus clientes. Neste caso, ele rejeitaria anúncios do AS 2 com prefixos do AS 1. Essa é uma estratégia bastante efetiva também para se evitar vazamentos de rota em que um cliente exporta acidentalmente rotas recebidas de um de seus provedores para seus demais provedores.

Por fim, sequestradores podem restringir o alcance de um sequestro propositalmente. Por exemplo, Milolidakis *et al.* [56] discutem e simulam vários cenários que os sequestradores podem explorar para realizar sequestros que não são visíveis pelos coletores BGP. Birge-lee *et al.* [8], por outro lado, explora o uso de comunidades BGP [27] para controlar a propagação do ataque e manter uma rota para a vítima.

Capítulo 4

Trabalhos Relacionados

Sequestros de prefixo têm sido estudados e analisados em diversos trabalhos ao longo do tempo [23, 39, 48, 72, 81, 84]. Shi *et al.* [75] mostraram pela primeira vez que os sequestros de prefixo, que até então só tinham sido considerados em teoria, ocorrem de fato na Internet. Desde então, vários pesquisadores têm se dedicado a identificar esses sequestros e caracterizar seus responsáveis. Este capítulo apresenta alguns dos principais trabalhos que buscam caracterizar os ASes responsáveis por sequestro bem como identificar sequestros no plano de dados, no plano de controle ou em ambos.

4.1 Caracterização de ASes Maliciosos

Identificar ASes responsáveis por sequestros de prefixo na Internet é desafiador devido à falta de segurança do protocolo BGP, que permite que um AS altere as informações de anúncios de rota, incluindo o *AS-path*. Em um sequestro do Tipo-0, é fácil identificar o sequestrador por ele estar no início do *AS-path* e não ser o proprietário do prefixo. Porém, em casos de *AS-path* forjados, a ocorrência e a identificação do AS responsável por um sequestro não são tão óbvias. Portanto, a caracterização do comportamento de um AS envolvido em sequestros ou com potencial malicioso para se tornar um sequestrador pode auxiliar na identificação de novos sequestros na Internet.

Shue *et al.* [76] apresentam uma análise extensiva sobre o comportamento de ASes maliciosos na Internet com o objetivo de caracterizar seus comportamentos. Inicialmente, eles selecionam os ASes maliciosos utilizando informações das dez listas de bloqueio mais comuns para *spam*, *phishing*, *malware* e *botnets*. A análise do comportamento desses ASes utilizou dados públicos de BGP do Route Views [60] e identificou quatro provedores com mais de 80% de seus endereços listados e 22 provedores com 100% de seus clientes envolvidos em atividades maliciosas (clientes com no mínimo 1% de seus endereços bloqueados). Em junho de 2009, 46,8% dos 31.263 ASes observados tinham pelo menos um endereço listado em uma das listas de bloqueio. O estudo revela que ASes maliciosos tendem a ficar ausentes da tabela de roteamento por curtos períodos e realizam mais mudanças de rotas do que ASes legítimos. Outras características dos ASes maliciosos incluem um maior número de endereços e mais conexões com outros ASes. O principal objetivo do trabalho é auxiliar na contratação de

provedores ou na escolha de rotas, confirmando a existência de ASes cuja principal função é facilitar ou permitir atividades maliciosas na Internet, mas sem se aprofundar nas atividades maliciosas específicas ou na classificação detalhada dos ASes.

Konte *et al.* [42] identificam uma categoria de Sistemas Autônomos que possuem estruturas de conexão que lhes permitem permanecer online mesmo quando estão envolvidos em atividades maliciosas, como hospedar serviços maliciosos ou sequestrar prefixos, e os denominam de *bulletproof* (*i.e., a prova de balas*). Para garantir sua disponibilidade contínua, um AS *bulletproof* muda de provedor frequentemente e anuncia prefixos com poucos endereços IP, permitindo a utilização de outros prefixos caso um deles seja bloqueado. Konte *et al.* identificam os ASes 12383, 50215, 31366, 44051 e 8287 como ASes que mantêm conexões com diversos ASes legítimos para ocultar as conexões dos ASes 50369, 42229, 50390, 12604, 49934, 47560, 44107 e 47821, que de fato hospedam sistemas maliciosos. O trabalho mostra que geralmente não se trata de apenas um ou outro AS envolvido em atividades maliciosas na Internet, mas sim de um conjunto de ASes, incluindo aqueles que servem de provedores para os ASes *bulletproofs*. Essa estratégia dificulta significativamente o banimento desses ASes da Internet.

Testart *et al.* [81] caracterizam ASes que sequestram prefixos regularmente e os denominam de *sequestradores seriais*. O trabalho faz uma análise extensa do comportamento desses ASes utilizando aprendizado de máquina com dados de coletores BGP de um período de cinco anos (de 1 de janeiro de 2014 a 31 de dezembro de 2018). Eles utilizam um modelo composto por 54 *features* para classificar um AS como legítimo ou sequestrador serial a partir de uma votação de 34 *Extra Trees* [36] treinadas com diferentes técnicas de amostragem. O conjunto de dados de treinamento inclui 217 ASes legítimos, selecionados entre participantes do MARNS (*Mutually Agreed Norms for Routing Security*) [52], e 23 ASes sequestradores seriais, identificados em fóruns e listas de discussão. O modelo treinado foi utilizado para classificar 19.103 ASes observados no período de cinco anos, identificando 900 possíveis sequestradores seriais. Alguns resultados foram falsos positivos, como a classificação equivocada de redes de proteção contra DDoS, pois elas possuem comportamentos semelhantes ao comportamento de um sequestrador, conforme discutido no Capítulo 3.

Com a identificação de alguns ASes que provavelmente são sequestradores seriais, foi possível criar um perfil desses sequestradores, destacando-se as seguintes características em comparação com os ASes legítimos:

- **Presença na Internet**: ASes sequestradores seriais não ficam visíveis 100% do tempo na tabela global de roteamento e apresentam períodos de ausência em que não originam prefixo algum. Por outro lado, ASes legítimos possuem presenças estáveis (quase 100% do tempo) e só ficam ausentes da tabela global de roteamento em situações de problema.

- **Anúncio de prefixos**: ASes sequestradores seriais tendem a anunciar uma grande quantidade de prefixos únicos. Porém, cada um desses prefixos é anunciado por um curto período de tempo. Em contrapartida, um AS legítimo geralmente mantém uma quantidade relativamente constante de prefixos anunciados por longos períodos, frequentemente superiores a um ano. Mudanças significativas na quantidade de

prefixos anunciados por ASes legítimos são observadas apenas em períodos breves e geralmente resultam da divulgação de prefixos mais específicos, possivelmente para implementar engenharia de tráfego.

Uma grande dificuldade relatada em [81], bem como em outros trabalhos que empregam técnicas de aprendizado de máquina supervisionado [39], é a construção do conjunto de dados de treinamento, pois sequestros de prefixo nem sempre são relatados. Geralmente, os proprietários dos prefixos sequestrados optam por não divulgar os problemas de segurança que enfrentaram com receio de prejudicar a reputação da entidade responsável pelo AS.

4.2 Identificação no Plano de Dados

A identificação de sequestros no plano de dados consiste em analisar os pacotes que trafegam pela rede para identificar sequestros, sem considerar as informações nas tabelas de roteamento BGP (*i.e.,* plano de controle). Uma abordagem comum nesse tipo de análise é observar as alterações no volume de tráfego de dados para um destino específico. Quando ocorre uma mudança anormal nesse volume, pode-se supor que há um sequestro de prefixo em andamento. Isso ocorre porque o volume de dados em direção à vítima diminui enquanto o volume com destino ao sequestrador aumenta. Essa ideia serve como base para o LDC (*Load Distribution Change*) [48].

A Figura 4.1 ilustra a ideia utilizada pelo LDC. Na situação ilustrada pela Figura 4.1(a), não há sequestro e todo o tráfego de dados destinado ao prefixo `200.0.0.0/24` é encaminhado para o AS 2, que então o repassa ao AS 1, proprietário do prefixo. A Figura 4.1(b) apresenta um exemplo de sequestro do prefixo `200.0.0.0/24` pelo AS 9. Nesse caso, todos os ASes cujas tabelas de roteamento foram afetadas pelo sequestro começam a redirecionar os pacotes destinados ao prefixo sequestrado para o AS 7, que por sua vez os encaminha ao AS 9. Com essa mudança no fluxo de dados, o AS 2 observa uma redução no volume de dados destinados ao AS 1, enquanto o AS 7 nota um aumento no tráfego com destino ao AS 9. O LDC identifica sequestros ao analisar essas variações de volume nos fluxos de dados. No entanto, essa análise não é trivial devido à variação normal do uso da Internet ao longo do dia e da semana. Para auxiliar na classificação, o LDC emprega um algoritmo de clusterização não supervisionado que considera não apenas o volume de tráfego, mas também sua variação ao longo do tempo. O LDC é idealmente implementado pelos provedores de acesso à Internet e oferece duas formas de implementação: uma instalação única (*Individual Mode*), como exemplificado pelo AS 2 na Figura 4.1(a), ou várias instalações que trabalham de forma cooperativa (*Cooperative Mode*), como exemplificado pelos ASes 2 e 7 na Figura 4.1(b). Com múltiplas instalações, torna-se mais fácil identificar a localização da vítima e do sequestrador.

Uma abordagem alternativa no plano de dados é a análise de RTT (*Round Trip Time*), que mede o tempo de ida e volta de um pacote. Em caso de sequestro, o RTT pode aumentar ou diminuir, dependendo da proximidade do sequestrador e das condições das redes percorridas pelo pacote. Bühler *et al.* [10] utilizam a análise de RTT para detectar

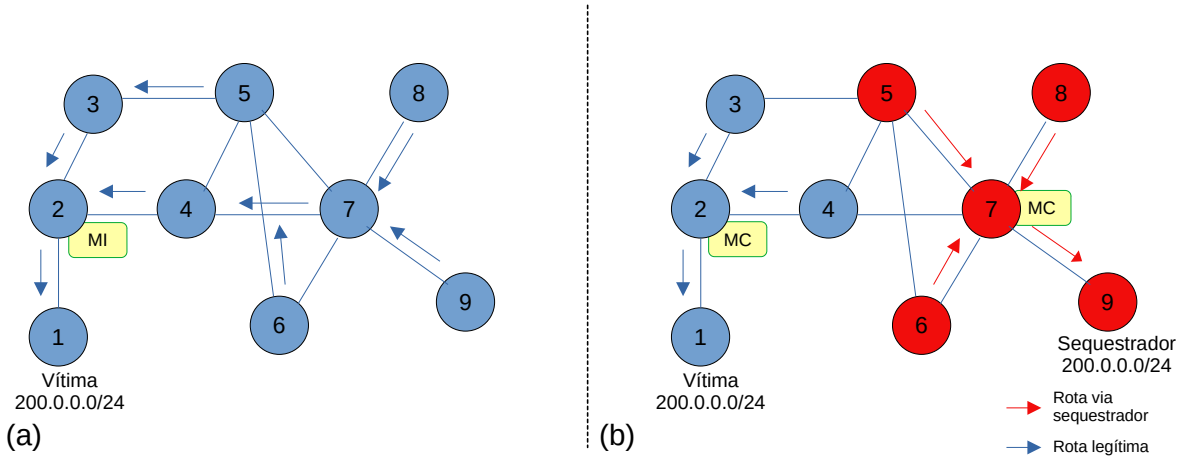


Figura 4.1: Fluxo de pacotes em situações normal e com implantação do LDC em modo individual (MI) no AS 2 (a) e com sequestro do prefixo 200.0.0.0/24 e com implantação do LDC em modo cooperativo (MC) nos ASes 2 e 7 (b).

sequestros, embora presumam que já tenham acesso aos valores de RTT e apenas citem alguns trabalhos que coletam essa informação de forma passiva. Uma outra possibilidade seria a análise ativa de RTT, com o uso de `ping`, que envia uma mensagem `echo request` e aguarda por uma resposta `echo reply` do ICMP (*Internet Control Message Protocol*) [64]. No entanto, essa abordagem ativa requer que o destino esteja ativo e que não haja configurações que bloqueiem as mensagens `echo request/reply`. Além disso, encontrar endereços de teste adequados para monitorar todos os destinos possíveis e lidar com a complexidade e o volume de tráfego necessários para testar uma grande quantidade de endereços são desafios adicionais [69, 82].

4.3 Identificação no Plano de Controle

A maioria dos trabalhos existentes na literatura para identificação de sequestros de prefixo utiliza informações do plano de controle [39, 43, 72–74, 81]. Uma razão para isso é a disponibilidade de coletores distribuídos pela Internet [60,66] que fornecem as informações essenciais para análises e avaliações experimentais. É importante notar que esses coletores possuem dados históricos, algo que é difícil de obter no plano de dados.

O Tipo-0, a princípio, é o tipo de sequestro de prefixo mais fácil de se identificar com dados do plano de controle, pois nele, múltiplos ASes anunciam o mesmo prefixo (MOAS). Entretanto, apenas a presença de MOAS não necessariamente indica um sequestro, já que outras situações podem levar a múltiplos ASes a anunciar o mesmo prefixo, como serviços de proteção contra DDoS e o uso de ASN privado em um cliente com múltiplos provedores. Neste último caso, os provedores removem o ASN privado e anunciam os prefixos do cliente como se fossem deles.

PHAS (*Prefix Hijack Alert System*) [43] é um sistema que tem como objetivo identificar e alertar os proprietários de prefixos sobre sequestros do Tipo-0. O sistema

recebe atualizações BGP dos coletores e verifica se algum prefixo é anunciado por uma nova origem. Os administradores de ASes interessados em usar o sistema devem se registrar e indicar os prefixos de interesse, além de fornecer um ou mais endereços de e-mail para receber notificações. Os autores sugerem usar endereços de e-mail não vinculados aos prefixos de interesse para evitar a possibilidade de os alertas não serem recebidos durante um sequestro. Um alerta é enviado sempre que uma nova origem é observada para um prefixo de interesse, deixando a análise sobre a natureza do evento — se é um sequestro ou um falso positivo — a cargo do usuário do sistema. Quando a divulgação do prefixo volta a ser feita apenas pelo ASN original, um novo e-mail é enviado para informar a normalização da situação.

Shapira e Shavitt [73] propõem um método de dois estágios para identificar sequestros de prefixo usando IA. O primeiro estágio cria *embeddings* (representações vetoriais) dos ASNs para representar suas características e relações, similar ao modelo *skip-gram* [55]. O modelo utiliza uma *shallow neural network* com uma camada de entrada contendo uma posição para cada AS observado (62.525 no período analisado), uma camada oculta totalmente conectada de tamanho 32, e uma saída determinada pelo tamanho da janela de tamanho dois (distância máxima entre o ASN de entrada e o ASN a ser predito na saída), resultando em uma saída de no máximo 4x62.525. No segundo estágio, uma LSTM (*Long Short Term Memory*) de cinco camadas processa a sequência de ASNs gerada pelo primeiro estágio, produzindo um vetor de tamanho 100. Esse vetor é então classificado por um único neurônio com uma função de ativação que rotula o *AS-path* como "verde" (rota padrão) ou "vermelho" (rota sequestrada). O modelo foi treinado com dados do Route Views [60] e rotas BGP marcadas como padrão ("verde") ou sequestradas ("vermelho") do BGPMon[1], atingindo uma acurácia de 99,98

Em outro trabalho, Shapira e Shavitt propõem o AP2Vec [74] para identificar sequestros de prefixo usando treinamento não supervisionado, eliminando a necessidade de rotas rotuladas. AP2Vec utiliza o modelo Word2Vec, mas com a abordagem CBOW (*Continuous Bag Of Words*) em vez de *skip-gram*. No CBOW, o modelo recebe palavras do contexto e tenta prever a palavra correspondente. O primeiro estágio gera *embeddings* dos ASNs e prefixos de endereço usando AP2Vec, com treinamento similar ao Doc2Vec [45], em que cada *AS-path* é tratado como uma sentença de um texto.

O segundo estágio gera os *embeddings* do prefixo e do *AS-path* para um primeiro anúncio observado em um ponto de observação (*vantage point* - VP). O primeiro anúncio observado é considerado o pai. Ao observar um anúncio diferente para o prefixo, ou um prefixo mais específico, os autores sugerem duas abordagens: a primeira analisa a similaridade entre os caminhos e os prefixos usando a similaridade cosseno entre os vetores do prefixo pai e o novo observado; a segunda analisa a distância entre os vetores dos caminhos pai e novo para identificar mudanças significativas nas estruturas dos ASes. A rota é considerada sequestrada se o valor exceder um limite aceitável de diferença entre as rotas. Após análises com amostras de diferentes anos e horários, as duas abordagens geraram entre dois e dezesseis eventos para cerca de dez mil mudanças de rotas de prefixos, um valor considerado razoável para ser tratado pelo time de resposta.

Artemis (*Automatic and Real-Time dEtection and MItigation System*) [72] é um

[1]https://bgpstream.com/about/

sistema de detecção e mitigação de sequestros de prefixo implantado localmente na rede de um AS que deseja monitorar seus prefixos. Ele utiliza informações fornecidas pelo administrador do AS e de coletores BGP para identificar sequestros do Tipo-0 e Tipo-1 principalmente. O administrador fornece informações como o prefixo monitorado, quais ASNs podem divulgá-lo e as conexões do AS local. Artemis compara essas informações com dados dos coletores e emite um alerta se detectar um ASN não autorizado divulgando o prefixo ou uma conexão não informada no *AS-path*. Para sequestros do Tipo-`X`, com `X` ≥ 2, Artemis mantém um histórico de 10 meses de ligações de ASes e emite um alarme caso detecte que uma nova conexão não foi registrada no histórico.

Sermpezis *et al.* [72] propõem o uso de Artemis para automatizar a mitigação de sequestros de prefixo. A ideia é que o sistema tenha acesso aos roteadores de borda e, ao detectar um sequestro, configure-os para divulgar o prefixo original sequestrado em vários prefixos mais específicos (/24 para IPv4 e /48 para IPv6), pois uma rota com prefixo mais específico terá prioridade no plano de dados. No entanto, se o sequestro envolver prefixos desses tamanhos, uma divulgação mais específica pode ser ineficaz devido às restrições da maioria dos ASes quanto ao tamanho dos prefixos aceitos. Essas restrições são uma prática comum para evitar que a tabela de roteamento cresça demais e sobrecarregue a memória dos equipamentos. Uma possível solução é usar um AS com ampla conectividade para divulgar o prefixo em várias regiões, forçando a escolha de rotas com *AS-path* mais curtos, e redirecionar o tráfego para o AS vítima, semelhante a uma rede anti-DDoS durante a proteção contra ataques.

4.4 Identificação em Ambos os Planos

Para reduzir os falsos positivos na identificação de sequestros no plano de controle, especialmente aqueles que causam perda de comunicação com o prefixo (*blackholing*), alguns trabalhos verificam se o prefixo está acessível no plano de dados após detectar um possível sequestro no plano de controle. Exemplos incluem os sistemas Argus [84] e Themis [65]. Shi *et al.* [75] apresentam uma análise do desempenho do Argus após um ano de operação e a partir dessa análise mostraram que o sequestro de prefixo existe na Internet, algo até então estudado somente na teoria. Também observaram que cerca de 20% dos sequestros duram menos de dez minutos e que podem poluir 90% da Internet em menos de dois minutos.

O sistema Argus [75, 84] é organizado em três módulos. O primeiro é o Módulo de Monitoramento de Anomalias (AMM, *Anomaly Monitoring Module*), que detecta três tipos de anomalias: divulgação de um prefixo por um AS não observado antes (origem anômala), ligação entre ASes não diretamente conectados (vizinho anômalo) e violação de política, quando um AS cliente aparece entre dois provedores (política anômala). O segundo módulo é o de Identificação de Sequestro (HIM, *Hijacking Identification Module*), e o terceiro é o Módulo de Coleta de IPs Vivos (LCM, *Live-IP Collection Module*), que obtém endereços IP para testes de conectividade. Quando o AMM identifica uma anomalia, ele aciona o HIM, que realiza testes de conexão usando `ping` por meio de sistemas remotos, chamados de "olhos do Argus". O LCM fornece os endereços para esses testes, incluindo o primeiro endereço válido de um prefixo (`.1`) e

endereços mapeados por `traceroute`. O Argus sinaliza um possível sequestro dependendo da quantidade de olhos que observaram a anomalia e nas respostas de `ping`.

O sistema Themis [65] aprimora o Argus visando focar na identificação de sequestros que causam MOAS, diferenciando MOAS legítimos de sequestros de prefixos. Esse trabalho também define as características que distinguem um sequestro de um anúncio legítimo, seleciona eventos confirmados de sequestro e treina um algoritmo de aprendizado de máquina (*Extra-trees*) para identificar sequestros não catalogados. Themis utiliza os módulos do Argus, mas inclui um classificador de MOAS antes do teste de conectividade realizado pelo HIM. Além disso, o Themis utiliza RPKI e informações dos RIRs para melhorar a identificação dos ASNs proprietários dos prefixos.

Capítulo 5

Caracterização de Sequestros de Prefixo

Sistemas de comunicação militares são essenciais para a segurança e soberania de uma nação. Enquanto que em cenários de comunicação crítica/sensível as entidades militares costumam ter suas próprias infraestruturas de comunicação, em cenários menos críticos a comunicação ocorre através da Internet. Exemplos incluem acesso a serviços de alistamento e gestão bem como acessos de membros das Forças Armadas que estão em missões diplomáticas fora de seu país. Por estarem expostos na Internet, estes sistemas tornam-se potenciais alvos de ataques cibernéticos, como o sequestro de prefixo.

Através do uso de um simulador de sequestro desenvolvido para este trabalho e apresentado na Seção 5.1, e tendo como prováveis vítimas ASes utilizados por militares (Seção 5.2), este capítulo apresenta os resultados obtidos nas simulações de sequestro quanto a contaminação dos demais ASes (Seção 5.3), a capacidade dos coletores em observar os sequestros (Seção 5.4), as características que tornam um AS mais resiliente ao sequestro (Seção 5.5) e o impacto que o uso de *prepend* na origem pode ter em relação à resiliência (Seção 5.6).

Aspectos éticos. As análises foram realizadas exclusivamente com dados públicos de roteamento. Os impactos de possíveis sequestros de prefixo foram avaliados por meio de simulações, sem qualquer interferência nas redes estudadas. Para preservar a privacidade e evitar a exposição de vulnerabilidades, as redes analisadas foram anonimizadas.

5.1 Simulação de Roteamento Interdomínio e Sequestros de Prefixo

Este trabalho simula cenários de sequestro de prefixo utilizando um simulador que desenvolvemos e disponibilizamos publicamente em [19]. As simulações utilizam dados do grafo de relacionamentos da CAIDA [13] para os meses de fevereiro, março e abril de 2024, abrangendo 76.421, 76.556 e 76.649 ASes, respectivamente. Os tipos de relacionamento entre os ASes (cliente-provedor (c2p) e *peer-to-peer* (p2p)) também são extraídos dos grafos da CAIDA. Apesar de não ser perfeito, o grafo da CAIDA é considerado a melhor fonte de relacionamento entre ASes na Internet e é frequentemente utilizado em estudos desta natureza [56].

O simulador aplica uma versão simplificada, mas realista, do processo de decisão do BGP para selecionar a melhor rota para um prefixo. Cada AS escolhe a melhor rota seguindo os critérios: *(i)* maior atributo de preferência local; *(ii)* em caso de empate, a rota com o menor *AS-path*; *(iii)* persistindo o empate, a primeira rota recebida. As preferências locais são determinadas pelo modelo Gao-Rexford [34], que reflete as relações comerciais entre ASes: rotas de clientes são preferidas sobre rotas de *peers*, que, por sua vez, têm prioridade sobre rotas de provedores. A propagação de rotas também segue o modelo Gao-Rexford: rotas recebidas de clientes são propagadas para todos os vizinhos, enquanto rotas de *peers* ou provedores são propagadas apenas para clientes.

Para melhor representar a realidade da Internet, as rotas analisadas são aquelas observadas nos ASes que exportam rotas para os coletores dos projetos RIPE-RIS [66] e Route Views [60], que são chamados de *pontos de observação* (*vantage points*). Os ASNs desses ASes foram obtidos diretamente nos sites desses projetos, totalizando 811, 813 e 815 ASes nas simulações de fevereiro, março e abril de 2024, respectivamente. Informações adicionais sobre os ASes, como descrição e país, foram coletadas do *CIDR Report* [63] em outubro de 2024. Além disso, na análise do impacto de um sequestro envolvendo um AS vítima que implementa ROA, a validação de origem com ROV é simulada apenas para ASes com *score* superior a 0,25, conforme os dados do estudo [47] na mesma data ou na mais próxima disponível dos dados utilizados. O *score* de cada AS foi obtido do mesmo estudo [47] e indica a chance do AS rejeitar um anúncio inválido. Por exemplo, um AS com *score* igual a um valida com ROV todos os anúncios que possuam ROA.

Apesar de utilizar informações obtidas da Internet, as simulações apresentam algumas limitações. Entre elas, pode-se destacar a falta de enlaces não observados pelos relacionamentos disponibilizados pela CAIDA, políticas distintas de roteamento quando há múltiplas conexões entre ASes e conexões via PTTs. Diferentes configurações utilizadas por ASes na Internet e que violam o modelo Gao-Rexford também não são consideradas.

5.2 Cenários das Simulações

Nas simulações, as vítimas de sequestro são ASes utilizados pelas Forças Armadas dos países do G20 ou ASes que anunciam prefixos pertencentes a essas entidades (*e.g.,* provedores de nuvem). Foram selecionados 29 ASes, distribuídos da seguinte forma: África do Sul (1), Alemanha (1), Arábia Saudita (1), Argentina (1), Austrália (2), Brasil (4), Canadá (1), China (1), Coreia (1), Estados Unidos (8), Índia (1), Indonésia (2), Itália (1), Japão (2), Rússia (1) e Turquia (1). Os ASNs desses ASes foram identificados por consultas em páginas de acesso público [30] com os termos “Exército”, “Marinha”, “Força Aérea” e “Departamento de Defesa”, no idioma de cada país. Além disso, foi realizada a resolução de nomes de páginas oficiais dessas Forças Armadas para endereços IP e a determinação do AS dono do endereço com a ferramenta `whois`.

Os ASes vítimas variam desde aqueles com apenas um vizinho até ASes com mais de 1.700 vizinhos. A Tabela 5.1 apresenta as características desses ASes com base nos relacionamentos registrados pela CAIDA para abril de 2024, incluindo o número total de

Tabela 5.1: Caracterização dos vizinhos dos ASes vítimas em abril de 2024.

	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO
Vizinhos	1	1	1	1	1	1	1	1	2	2	2	2	2	2	4
Países	1	1	1	1	1	1	1	1	1	1	2	1	2	1	1
Continentes	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1
Clientes	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Peers	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Provedores	0	1	1	1	1	1	1	1	2	2	2	2	2	2	4

	AP	AQ	AR	AS	AT	AU	AV	AW	AX	AY	AZ	BA	BB	BC	
Vizinhos	5	7	7	7	16	17	31	32	40	73	276	597	1518	1765	
Países	1	1	1	1	1	1	1	10	12	3	46	31	92	142	
Continentes	1	1	1	1	1	1	1	3	4	1	5	5	5	5	
Clientes	3	4	6	1	0	15	30	0	0	70	3	389	1462	660	
Peers	0	0	0	4	15	0	0	30	25	0	226	200	56	848	
Provedores	2	3	1	2	1	2	1	2	15	3	47	8	0	257	

vizinhos, países e continentes envolvidos e vizinhos por tipo de relacionamento.[1] Entre as vítimas está o AS BC, vinculado a um grande serviço de hospedagem que também abriga sistemas de uma das Forças Armadas do G20. Essa relação foi confirmada pela resolução de nomes de páginas oficiais, cujo endereço IP é propriedade do AS BC.

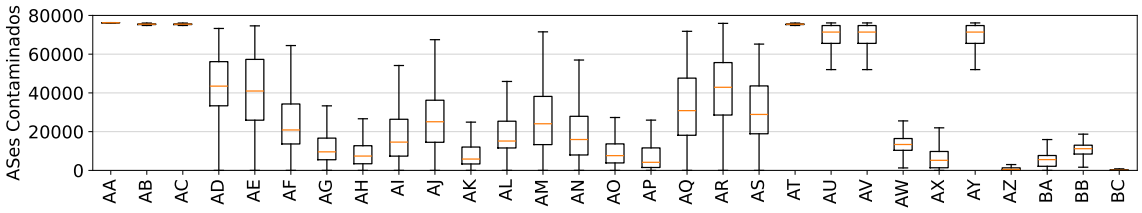
Os sequestradores foram selecionados aleatoriamente, totalizando 600 ASes por data de simulação, distribuídos em quatro grupos baseados no número de vizinhos para garantir a inclusão de ASes de diferentes tamanhos. A seleção incluiu 150 ASes de cada grupo, com os sequestradores permanecendo os mesmos para todas as simulações de uma mesma data. A composição dos grupos e o número de ASes em cada um nos primeiros dias de fevereiro, março e abril de 2024 são as seguintes:

- Grupo 0 (G0): ASes com dois vizinhos (22.675, 22.782 e 22.530 ASes);
- Grupo 1 (G1): ASes com três vizinhos (7.274, 7.345 e 7.444 ASes);
- Grupo 2 (G2): ASes com quatro a dez (8.721, 8.661 e 8.807 ASes);
- Grupo 3 (G3): ASes com 11 vizinhos ou mais (9.801, 9.788 e 9.904 ASes).

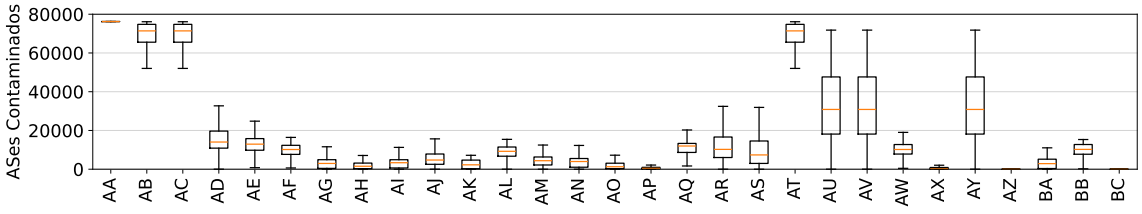
Os ASes com apenas um vizinho (*stub*) não foram selecionados como sequestradores por não se observar sequestros maliciosos noticiados tendo como sequestrador um AS deste tipo. Além disso, um AS *stub* é incapaz de realizar sequestros de interceptação de tráfego por não possuir caminho alternativo para repasse dos dados para a vítima.

Por fim, foram realizados três tipos de simulações para cada AS vítima e sequestrador: sequestros Tipo-0, Tipo-1 e Tipo-0 com ROV, simulando a validação de origem por RPKI. As simulações utilizaram dados do primeiro dia de fevereiro, março e abril de 2024 e todos os sequestros são realizados para prefixos com o mesmo comprimento dos da vítima, pois anúncios de prefixos mais ou menos específicos se propagariam por toda a Internet, inviabilizando análises específicas, como a visibilidade nos coletores.

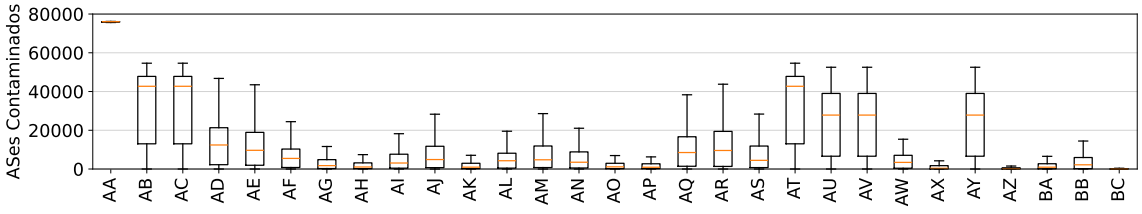
[1] Um AS sem provedor é Tier-1 e o outro possui ligação com um *sibling* que lhe provê trânsito.



(a) Sequestros Tipo-0



(b) Sequestros Tipo-1



(c) Sequestros Tipo-0 com ROV

Figura 5.1: Quantidade de ASes contaminados para os 600 sequestros simulados por vítima para os dados referentes ao mês de abril de 2024.

5.3 Contaminação dos ASes pelo Sequestro

A primeira análise avaliou a capacidade de contaminação das tabelas de roteamento dos ASes da Internet por tipos de sequestro. A Figura 5.1 mostra o número de ASes afetados por sequestros de Tipo-0 (a), Tipo-1 (b) e Tipo-0 com ROV ativo (c). Os ASes estão organizados na figura pela quantidade de vizinhos, conforme a ordem da Tabela 5.1. Os resultados apresentados referem-se a abril de 2024, com padrões similares observados em fevereiro e março. O sequestro Tipo-0 exibiu a maior capacidade de contaminação, como esperado, contaminando uma média de 31.117 (40,7%) ASes por sequestro. Sequestros Tipo-1 contaminaram menos, pois o *AS-path* maior no anúncio inicial reduz sua preferência quando o tamanho do *AS-path* é usado como critério de escolha, contaminando em média 17.094 (22,3%) ASes por sequestro. O sequestro Tipo-0 com ROV teve o menor impacto, devido ao descarte da rota sequestrada por ASes que realizam validação de origem, contaminando em média 11.398 (14,9%) ASes por sequestro. Além do tipo de sequestro, fatores como o número de vizinhos do AS vítima, o tipo de conexão e a dispersão geográfica dos vizinhos influenciaram significativamente o alcance da contaminação.

Os ASes AA, AB, AC e AT foram consistentemente os mais afetados pelos sequestros. Suas características comuns incluem poucas conexões com provedores e vizinhos

concentrados em um único país. Por outro lado, os ASes AZ e BC foram os menos impactados, devido à grande quantidade de vizinhos, principalmente provedores, e à diversidade geográfica desses vizinhos.

Outro fator relevante para o alcance de um sequestro é a conectividade do sequestrador. A Tabela 5.2 mostra que os sequestros provenientes de ASes do Grupo 3 (G3) apresentaram o maior alcance, pois esses sequestradores anunciaram o prefixo sequestrado por múltiplos caminhos em diferentes pontos da Internet.

Tabela 5.2: Média (μ) de contaminação dos sequestros, em valores absolutos e percentuais, por data, tipo de sequestro, grupo do sequestrador e geral.

Data	**Tipo-0**					**Tipo-1**					**Tipo-0+ROV**				
	μ	**G0**	**G1**	**G2**	**G3**	μ	**G0**	**G1**	**G2**	**G3**	μ	**G0**	**G1**	**G2**	**G3**
Fev	30457	27128	29015	31267	34417	16549	14451	15677	17087	18979	10565	7252	9563	11526	13917
2024	39,9%	35,5%	38,0%	40,9%	45,0%	21,7%	18,9%	20,5%	22,4%	24,8%	13,8%	9,5%	12,5%	15,1%	18,2%
Mar	31517	27675	29611	32529	36254	17279	15158	16121	17880	19957	11344	8418	9432	12616	14909
2024	41,2%	36,2%	38,7%	42,5%	47,4%	22,6%	19,8%	21,1%	26,4%	26,1%	14,8%	11,0%	12,3%	16,5%	19,5%
Abr	31376	27091	29992	32183	36239	17453	15058	16562	17860	20333	12285	8113	10802	13403	16824
2024	40,9%	35,3%	39,1%	42,0%	47,3%	22,8%	19,6%	21,6%	23,3%	26,5%	16,0%	10,6%	14,1%	17,5%	21,9%
μ	31117	27298	29539	31993	35636	17094	14889	16120	17609	19756	11398	7928	9932	12515	15217
Global	40,7%	35,7%	38,6%	41,8%	46,6%	22,3%	19,5%	21,1%	23,0%	25,8%	14,9%	10,4%	13,0%	16,3%	19,9%

5.4 Visibilidade dos Sequestros nos Coletores Públicos de Rota

As simulações permitiram analisar a quantidade de sequestros que não alcançaram nenhum ponto de observação e, consequentemente, não teriam visibilidade nos coletores públicos. Os sequestros de Tipo-0 com ROV foram os que mais produziram sequestros não observados, o que ocorre por, normalmente, contaminarem menos ASes na Internet. A simulação com dados de 01/02/2024 do sequestro de Tipo-0 com ROV foi a que gerou o maior número de sequestros não visualizados, totalizando 3.255 (18,71%) de 17.400 sequestros simulados (150 sequestros por cada um dos quatro grupos e 29 vítimas). O sequestro de maior impacto não detectado contaminou as tabelas de roteamento de 2.402 ASes de um total de 76.556, representando 3,14% da Internet simulada com dados de 01/03/2024. A maior média de ASes contaminados não observados foi de 135, registrada para sequestros Tipo-0 e com dados de 01/03/2024. Os resultados detalhados por data e por tipo de sequestro estão na Tabela 5.3.

Segundo [72], sequestros que contaminam mais de 2% da Internet são observados nos coletores. No entanto, algumas simulações deste trabalho mostraram sequestros que afetaram mais de 3% dos ASes sem serem observados. Essa discrepância pode ser atribuída ao aumento das conexões entre ASes nos últimos anos. Comparando os dados de número de ASes e relacionamentos entre ASes, utilizados em [72], com os dados mais recentes deste trabalho de abril de 2024, observa-se que, enquanto o número de ASes aumentou em 34,4% (de 57.027 para 76.649), o número de ASes com mais de mil vizinhos aumentou em 59,3% (de 91 para 145) e o maior número de vizinhos de um AS aumentou em 50,3% (de 6.488 para 9.754). Além disso, a seleção de sequestradores com alta conectividade (Grupo

Tabela 5.3: Total de sequestros não observados por tipo e data, incluindo o sequestro com maior contaminação (percentual de ASes da Internet afetados) e a média de ASes contaminados.

Dados da Simulação	Tipo	Não Observados	Maior Contaminação	Média de Contaminados
01-02-2024	0	336 (1,93%)	2170 (2,84%)	125
01-02-2024	1	1089 (6,26%)	846 (1,11%)	47
01-02-2024	0 + ROV	3255 (18,71%)	1663 (2,18%)	38
01-03-2024	0	313 (1,80%)	2402 (3.14%)	135
01-03-2024	1	972 (5,59%)	1354 (1,77%)	32
01-03-2024	0 + ROV	3038 (17,46%)	1739 (2,27%)	29
01-04-2024	0	326 (1,87%)	842 (1,10%)	91
01-04-2024	1	926 (5,32%)	900 (1,17%)	27
01-04-2024	0 + ROV	2869 (16,49%)	925 (1,21%)	27

3) pode ter contribuído para a maior contaminação sem visibilidade nos coletores. Essa diferença nos resultados também reflete limitações em [72], que não detalha os critérios de seleção de vítimas e sequestradores.

5.5 Resiliência das Vítimas ao Sequestro de Prefixo

A resiliência de um AS a sequestros de prefixo mede a proporção de tráfego destinado à vítima que permanece inalterada durante o ataque. Em termos práticos, indica a porcentagem de ASes na Internet que provavelmente não será afetada quando um determinado AS for a vítima. Conforme [9], a resiliência de um AS (R_{AS}) é calculada como a média das resiliências individuais de cada sequestro (R_s), dividindo o somatório das resiliências pelos sequestros simulados (Q_{ss}). A resiliência de um AS a um sequestro específico (R_s) é definida como:

$$R_s = \frac{\text{ASes não contaminados} - 2}{\text{Total de ASes} - 2} = \frac{ASes_{nc} - 2}{T_{ASes} - 2},$$

o valor 2 é subtraído para excluir o sequestrador e a vítima do cálculo. A resiliência total, portanto, é dada por $R_{AS} = \frac{\sum R_s}{Q_{ss}}$. Neste trabalho, a resiliência foi calculada utilizando os resultados dos três tipos de sequestros simulados: Tipo-0, Tipo-1, Tipo-0 com ROV, conforme descrito na Seção 5.1. A Figura 5.2 apresenta a resiliência por AS para as simulações com os dados de 01/04/2024.

Atualmente, não existe uma técnica amplamente implementada capaz de bloquear a propagação de sequestros com *AS-path* forjado. No entanto, como mostrado na Figura 5.2, os ASes demonstram maior resiliência a sequestros de Tipo-1 em comparação aos de Tipo-0. Essa maior resiliência ocorre porque o *AS-path* em um sequestro de Tipo-1 é mais longo no início pela adição do ASN da vítima como origem, o que o coloca em desvantagem no critério de desempate baseado no comprimento do *AS-path*.

O uso correto de ROA+ROV pode aumentar significativamente a resiliência de um AS contra sequestros de Tipo-0, como observado em [51]. Por exemplo, a resiliência dos ASes AU, AV, e AY aumenta de 0,1271 para 0,6817, um incremento de 0,5546. O menor aumento foi observado no AS BC, com resiliência subindo de 0,9948 para 0,9978. Entre os 29 ASes analisados, apenas 13 possuem ROA configurada, e apenas oito a aplicam

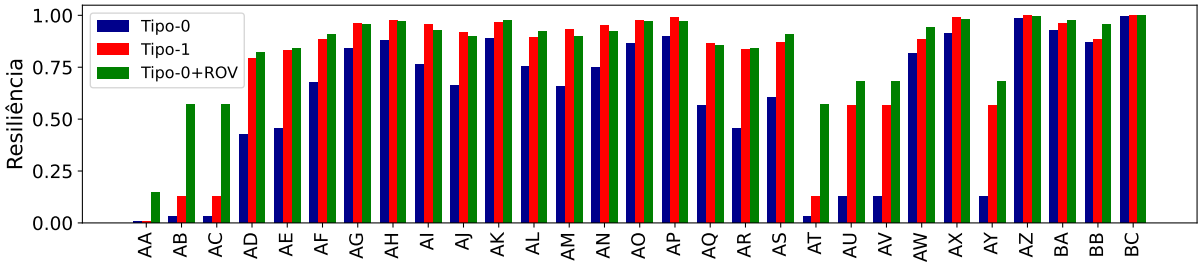


Figura 5.2: Resiliência de cada AS de acordo com o tipo de sequestro e com simulações realizadas com dados de abril de 2024.

para todos os seus prefixos. Em 01/04/2024, cerca de 49% dos prefixos IPv4 e 51,8% dos IPv6 eram validados pela ROV [59], indicando a necessidade de maior adoção. Com base na resiliência dos ASes (Figura 5.2) e nas características dos seus vizinhos (Tabela 5.1), pode-se concluir que:

- ASes bem conectados: Aqueles com mais provedores e maior distribuição geográfica são mais resilientes, como AZ, BA, BB e BC;
- ASes com poucos provedores: Mesmo com muitos enlaces *peer-to-peer*, tendem a ser menos resilientes, como o AS AT;
- ASes com baixa conectividade: Dependem da conectividade dos vizinhos, especialmente provedores, para maior resiliência. Por exemplo, o AS AH, com apenas um provedor, alcança uma resiliência de 0,8792 para o sequestro Tipo-0 porque seu provedor possui 2.992 vizinhos em 112 países diferentes.

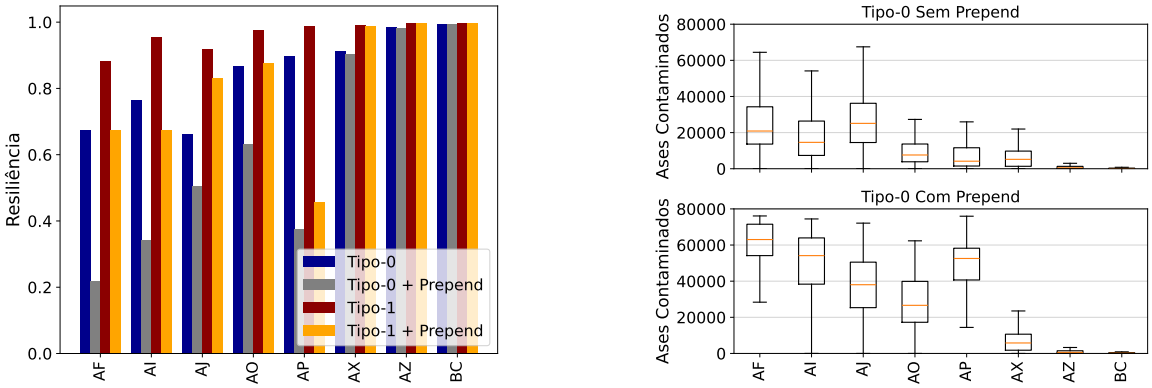
5.6 Impacto do *Prepend* no Sequestro

Os resultados apresentados nas Seções 5.3, 5.4 e 5.5 basearam-se em simulações de rotas propagadas sem técnicas de engenharia de tráfego. Analisando as RIBs disponibilizadas pelos coletores em 01/04/2024, observou-se o uso significativo de *prepend* nas rotas anunciadas pelos ASes vítimas. Esta técnica consiste em aumentar artificialmente o comprimento do *AS-path* repetindo uma ou mais vezes o ASN do AS realizando *prepend* [57]. O uso desta técnica serve para priorizar ou despriorizar determinados vizinhos na entrada do tráfego, mas pode facilitar sequestros de prefixo, já que anúncios legítimos com *prepend* tendem a ser preteridos devido ao maior *AS-path* [53].

Em 01/04/2024, todos os ASes militares da Tabela 5.1 tinham anúncios com *prepend*, e alguns aplicavam a técnica na origem. As rotas do AS AD foram as que mais apresentaram *prepend*, com 79,76% dos anúncios. Entre os ASes que implementaram *prepend* na origem, destacam-se AO (68,82% dos anúncios), seguido por AJ, AZ, BC, AF, AX, AI, e AP, com percentuais variando de 2,27% a 62,96%.

Para avaliar o impacto de *prepend* na resiliência e contaminação, foram analisados os ASes que aplicaram *prepend* na origem, o número de vezes que o ASN estava repetido e

os vizinhos afetados. Os dados das RIBs de 01/04/2024 foram usados nas simulações, e os resultados comparando resiliência com e sem *prepend* estão nas Figuras 5.3a e 5.3b.



(a) (b)

Figura 5.3: Em (a) a comparação dos valores de resiliência obtidos para os sequestros com e sem o uso de *prepend*. Em (b) na parte superior estão os valores de ASes contaminados para o sequestro Tipo-0 quando não há o uso de *prepend*, e na parte inferior quando há o uso do *prepend*.

Os resultados mostram que o uso de *prepend* pode reduzir significativamente a resiliência. Por exemplo, para o AS AP, a resiliência caiu de 0,8985 para 0,3770 (Tipo-0) e de 0,9899 para 0,4564 (Tipo-1). No entanto, ASes com vizinhos geograficamente distribuídos, como AX, AZ, e BC, foram menos impactados. Entre eles, a maior queda de resiliência foi no AS AX, com reduções de apenas 0,0077 (Tipo-0) e 0,0012 (Tipo-1), demonstrando que conexões diversificadas ajudam a manter a resiliência contra sequestros.

Capítulo 6

Detecção de Sequestros de Origem Forjada

Recentemente, sequestradores têm utilizado *AS-paths* forjados para evitar a detecção pela ROV (RPKI), causando grandes prejuízos aos ASes proprietários dos prefixos, especialmente aqueles envolvidos em transações com criptomoedas. Esses sequestros ocorrem quando um atacante manipula o *AS-path* de um anúncio para incluir o AS de origem do prefixo, fazendo o anúncio parecer legítimo. Esse tipo de sequestro ocorreu nos casos dos sequestros que resultaram no roubo de 1,9 milhão de dólares da plataforma KlaySwap [78] e de aproximadamente 235 mil dólares da Celer Bridge [41]. No incidente da Celer Bridge, houve dois sequestros: o primeiro para obter um certificado digital e o segundo para desviar as criptomoedas.

Este capítulo descreve e avalia o modelo de aprendizado de máquina (AM) do sistema DFOH (*Detect Forged-Origen Hijack*) [39], que foi desenvolvido recentemente para detectar sequestros de origem forjada e é considerado o estado-da-arte no momento. Além disso, todo o código fonte do sistema está disponível, o que permite a reprodução dos experimentos descritos em [39] e a realização de novos experimentos para validar e aprimorar o modelo proposto.

6.1 Modelo de AM do DFOH

O objetivo do DFOH [39] é detectar anúncios em que a conexão entre dois ASes consecutivos no *AS-path* seja forjada. Para isso, o sistema utiliza um modelo de aprendizado de máquina do tipo caixa-preta baseado em uma floresta aleatória (*random forest*) com 28 *features* calculadas a partir de diversas fontes públicas de informação de roteamento, como coletores de rota [60, 66], bases de relacionamento da CAIDA [12], informações de registros regionais [54] e peeringDB [62]. Em [39], os autores apresentam o DFOH como o sistema mais avançado e atual para detecção de sequestros de prefixo com origem forjada e mostram que ele atinge altas taxas de precisão e *recall*, sem gerar muitos alarmes falsos.

Tabela 6.1: *Features* utilizadas no modelo de floresta aleatória do DFOH.

Categoria			***Feature***	**Informação utilizada para computar a *feature***
Topológicas	Por AS	Centralidade	*degree_ centrality_ as1*	Fração dos ASes existentes conectados no AS1
			degree_ centrality_ as2	Fração dos ASes existentes conectados no AS2
			closeness_ centrality_ as1	Comprimento médio do caminho mais curto de todos os ASes até o AS1
			closeness_ centrality_ as2	Comprimento médio do caminho mais curto de todos os ASes até o AS2
			harmonic_ centrality_ as1	Média harmônica dos caminhos mais curtos do AS1 até os demais ASes
			harmonic_ centrality_ as2	Média harmônica dos caminhos mais curtos do AS2 até os demais ASes
		Vizinhos	*average_ neighbord_ degree_ as1*	Valor médio da quantidade de ASes conectados aos vizinhos do AS1
			average_ neighbord_ degree_ as2	Valor médio da quantidade de ASes conectados aos vizinhos do AS2
			eccentricity_ as1	Distância máxima do AS1 até os demais ASes
			eccentricity_ as2	Distância máxima do AS2 até os demais ASes
		P. Topo.	*triangles_ as1*	Número de ligações em triângulo que envolvem o AS1
			triangles_ as2	Número de ligações em triângulo que envolvem o AS2
			clustering_ as1	Fração dos possíveis triângulos existentes que incluem o AS1
			clustering_ as2	Fração dos possíveis triângulos existentes que incluem o AS2
	Par de ASes	Prox.	*jaccard*	Similaridade de Jaccard entre os vizinhos dos AS1 e AS2
			adamic_ adar	Proximidade de AS1 e AS2 baseada nos vizinhos compartilhados
			preferencial_ attachement	Probabilidade do AS1 e AS2 se conectar com base na quantidade de vizinhos
		Dist.	*shortest_ path*	Comprimento do caminho mais curto entre AS1 e AS2 com base no histórico
Peering			*country_ dist*	Países onde os ASes vizinhos são registrados
			ixp_ dist	PTT (IXP) onde os ASes vizinhos possuem conexões
			facility_ fac_ dist	Locais onde os ASes vizinhos possuem instalações
			facility_ cities_ dist	Cidades onde os ASes vizinhos possuem instalações
			facility_ country_ dist	Países onde os ASes vizinhos possuem instalações
AS-path			*degree*	Quantidade de ASes vizinhos
			cone	Quantidade de clientes diretos e indiretos
			cone_ degree	Valor de cone e de degree
Bidirec.			*bidi*	Se o enlace foi observado (1) ou não (0) em ambos os sentidos
			nb_ vps	Quantidade de vizinhos conectados em fornecedores de informações para os coletores

6.1.1 *Features* do Modelo

As *features* do modelo de DFOH, mostradas na Tabela 6.1, são divididas em quatro categorias: *topológicas, peering, padrão do AS-path e bidirecionalidade*. A tabela apresenta uma breve descrição de cada *feature*, mas a descrição detalhada pode ser encontrada em [39]. Para calcular as *features* topológicas, DFOH constrói um grafo direcionado representando as conexões entre os ASes usando anúncios BGP coletados durante os 300 dias anteriores ao treinamento. Como DFOH não utiliza dados de todos os coletores públicos disponíveis, ele complementa o grafo topológico com os *AS-paths* utilizados pela CAIDA para construção das informações de relacionamento entre ASes fornecidas em [12]. As *features* de *peering* são calculadas a partir de dados do PeeringDB [62]. Para calcular as *features* de padrão do *AS-path*, DFOH treina uma floresta aleatória utilizando sequências de grau de AS e tamanhos de cone de clientes a partir dos ASes presentes nos *AS-paths*. O cone de clientes é o conjunto de ASes clientes diretos e indiretos de um determinado AS. O grau de um AS é calculado a partir do grafo topológico e os tamanhos de cone de clientes são obtidos do ASRank [14]. A *feature* de bidirecionalidade indica se uma ligação entre dois ASes é bidirecional. Ela é construída com informações do grafo topológico e complementada com informações dos RIRs [54]. Por último, o valor da *feature nb_ vps*

é calculado com informações dos vizinhos observados no grafo topológico e a relação dos ASes que fornecem informações aos coletores BGP públicos [60, 66].

6.1.2 Amostragem e Treinamento do modelo

Para treinar o modelo, DFOH gera sinteticamente mil amostras de enlaces legítimos e mil de enlaces suspeitos envolvendo grupos de ASes determinados pelo algoritmo de clusterização `K-means` [1] e com dados dos 60 dias anteriores ao treinamento. A ideia é criar um conjunto de amostras representativas para ambas as classes que contenham ASes de todos os tipos (*i.e.,* Tier-1, *stub*, trânsito e *multi-homed*), pois uma geração totalmente aleatória favoreceria enlaces de ASes do tipo *stub* por serem em maior quantidade na Internet, cerca de 31% dos ASes em primeiro de dezembro de 2022 e de 2023. Note que o treinamento é realizado com dados dos 60 dias anteriores enquanto o grafo topológico é gerado com dados dos 300 dias anteriores. O período maior para a construção do grafo topológico visa capturar relações mais estáveis entre os ASes.

DFOH opera continuamente e retreina o modelo de floresta aleatória diariamente para analisar todos os novos enlaces surgidos por dia. Os novos enlaces classificados como legítimos podem ser usados para retreinar o modelo nos dias subsequentes, enquanto os classificados como forjados são removidos do grafo topológico por 30 dias, permitindo que os operadores de rede verifiquem possíveis sequestros envolvendo os ASes do enlace.

6.2 Desvendando o Funcionamento de DFOH com XAI

Nos últimos anos, aprendizado de máquina tem sido utilizado para resolver diversos problemas, empregando modelos cada vez mais complexos baseados em redes neurais profundas, florestas aleatórias e LLMs (*Large-Language Models*). Esses modelos são considerados caixas-pretas (*black boxes*) e geralmente são preferidos aos modelos interpretáveis, como árvores de decisão, devido ao seu desempenho superior. Entretanto, apesar da atenção que recebem, operadores de rede ainda hesitam em adotar modelos caixas-pretas em situações críticas, como bloqueio de tráfego, por não entenderem como as decisões são tomadas [40, 83]. Além disso, esses modelos complexos exigem uma quantidade maior de *features* que podem depender de informações de fontes nem sempre confiáveis e demandam maior tempo de processamento para computá-las.

Diversos esforços na área de Inteligência Artificial Explicável (XAI) propõem métodos para explicar ou examinar a validade das classificações geradas por modelos caixa-preta e determinar a relevância das *features* no processo de decisão. As abordagens de XAI podem ser divididas em dois grupos principais: *explicabilidade local* e *explicabilidade global* . Os métodos de explicabilidade local procuram identificar a importância de cada *feature* em uma decisão específica [68]. Por outro lado, a explicabilidade global visa explicar como o modelo caixa-preta toma decisões de forma geral e não em casos individuais. Geralmente, os métodos de explicabilidade global extraem um modelo caixa-branca (como uma árvore de decisão ou regras de decisão) que pode ser compreendido por um especialista no domínio do problema [40, 44].

A ferramenta Trustee [40], destinada à explicabilidade global, gera árvores de decisão de alta fidelidade para explicar o funcionamento de um modelo caixa-preta. Cada árvore de decisão é gerada utilizando aprendizado por imitação (*imitation learning*) [44] usando uma dinâmica professor-aluno em que um conjunto de amostras é passado para o modelo caixa-preta classificar. As amostras classificadas pelo modelo caixa-preta (professor) são usadas para treinar a árvore de decisão (aluno). O objetivo é gerar árvores de decisão que imitam o funcionamento do modelo caixa-preta. Assim, um especialista pode analisar as árvores de decisão para verificar se as decisões tomadas pelo modelo estão de acordo com o conhecimento existente no domínio do problema. Portanto, a árvore de decisão gerada por Trustee atua como um modelo substituto compreensível para o especialista. Trustee já foi utilizada para identificar problemas em modelos caixas-pretas em diferentes domínios de aplicação, especialmente na área de segurança de redes [6, 40].

O algoritmo de geração de árvores de decisão de Trustee (Algoritmo 1 em [40]) é agnóstico ao modelo caixa-preta sendo avaliado, ou seja, ele não utiliza parâmetros específicos do modelo. Entretanto, ele possui quatro parâmetros que influenciam na estabilidade da explicação (*i.e.,* árvore de decisão) gerada: número de amostras para treinar uma árvore de decisão (M), número de iterações do laço interno (N), número de iterações do laço externo (S), e número de ramos a ser considerado nas árvores com poda (k). O parâmetro M é especificado como uma fração do número total de amostras existentes no dataset de treinamento. Inicialmente, o algoritmo gera um dataset em que cada amostra é gerada a partir da classificação de uma amostra do dataset original pelo modelo caixa-preta. Em seguida, o algoritmo treina N árvores de decisão com M amostras selecionadas aleatoriamente do dataset gerado pelo modelo caixa-preta (laço interno) e seleciona a árvore de decisão com maior fidelidade ao modelo caixa-preta. Durante cada uma das N iterações desse passo, o algoritmo estende o dataset de treinamento com amostras que a árvore de decisão classifica de maneira diferente da classificação dada pelo modelo caixa-preta. A ideia é aumentar a quantidade de amostras que geram classificações divergentes, mas corrigindo-as com a classificação do modelo caixa-preta. Com isso, as novas árvores de decisão geradas tendem a concordar mais frequentemente com o modelo caixa-preta e assim melhorar suas fidelidades. O laço interno é repetido S vezes (laço externo) para se evitar explicações incorretas decorrentes de amostras ruins, gerando portanto S árvores de decisão. Cada uma dessas S árvores é podada para conter no máximo k ramos. Finalmente, o algoritmo escolhe a árvore de decisão com poda que possui a maior concordância média com as demais $S-1$ árvores do conjunto gerado pelo laço externo. Note que o algoritmo gera um total de $N \times S$ árvores de decisão para escolher aquela que fornece a explicação mais estável do modelo caixa-preta.

6.2.1 Abrindo a Caixa-Preta de DFOH com Trustee

DFOH utiliza uma floresta aleatória com 28 *features* calculadas a partir de várias fontes, como coletores BGP, registros regionais e PeeringDB. O modelo é bastante complexo, pois várias *features* não possuem interpretações diretas, sendo derivadas de cálculos que capturam o posicionamento de um AS na topologia da Internet ou suas interações com

outros ASes. A situação se complica ainda mais porque cada *feature* de *AS-path* (veja Tabela 6.1) é calculada a partir de outro modelo de floresta aleatória, tornando a compreensão do modelo inviável até mesmo para especialistas no domínio do problema. Entretanto, ao utilizar a ferramenta Trustee para explicar o modelo utilizado por DFOH, ela gera uma árvore de decisão, mostrada na Figura 6.1, com apenas cinco *features*, mas com fidelidade de 95% ao modelo de floresta aleatória original. Isso indica que, apesar de sua complexidade, o modelo não utiliza todas as *features* em seu processo de decisão e pode ser substituído por um modelo mais simples. As seções a seguir detalham a análise realizada sobre o modelo de DFOH, explicando como ele pode ser modificado para eliminar certas *features* sem afetar o seu desempenho. Inicialmente, são descritos os conjuntos de dados utilizados na análise e em seguida os resultados obtidos.

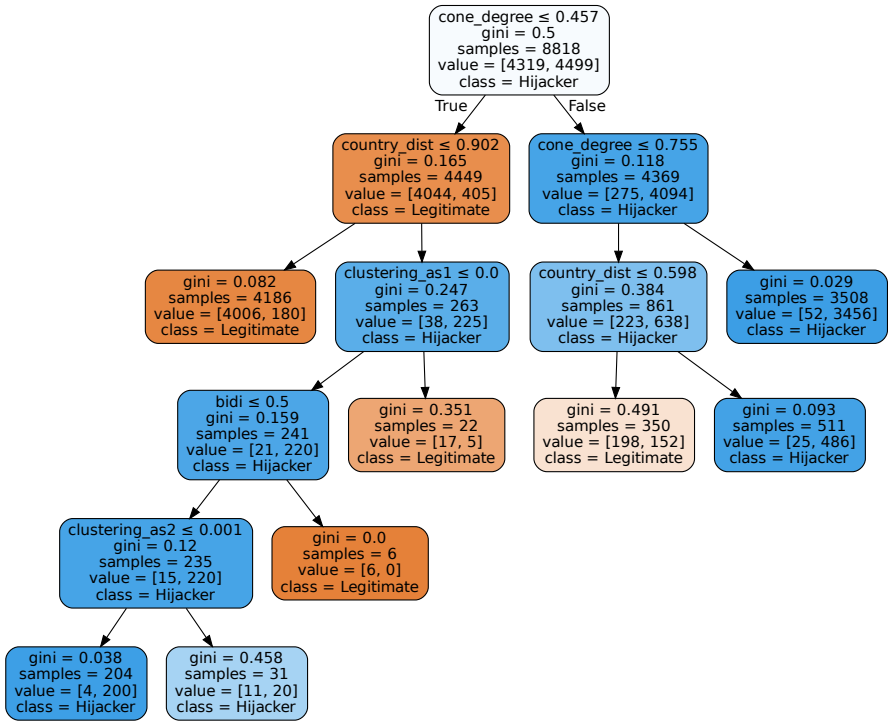


Figura 6.1: Árvore de decisão gerada por TRUSTEE para o modelo de DFOH.

6.3 Ambiente de Análise e Conjuntos de Dados

O código fonte de DFOH está disponível publicamente, o que facilita a reprodução dos resultados apresentados em [39] e a experimentação com mudanças no modelo. Durante a execução de DFOH, diversos arquivos são buscados de bases de dados públicas da Internet para cálculo das *features* do modelo de floresta aleatória. Para evitar que o tempo de download desses arquivos interferisse nas medições de tempo que são apresentadas na

Seção 6.4, todos os arquivos necessários para cálculos das *features*, treinamento do modelo e inferências foram copiados e salvos localmente no servidor usado nos experimentos. O servidor é uma máquina virtual com 10 núcleos de processamento Intel Xeon E5-2670 de 2,60 GHz isolados exclusivamente para as medições de tempo de execução e 124 GB de memória RAM rodando o sistema operacional Linux Ubuntu Server 20.04. O código fonte desenvolvido para a avaliação está disponível em [17].

Para avaliar o modelo de DFOH e suas modificações, foram utilizados conjuntos de dados gerados para dois períodos distintos em dezembro de 2022 e dezembro de 2023 com as seguintes características:

- **Dias analisados**: para cada período, foram analisados 20 dias, de 1 a 20 de dezembro, totalizando 40 dias;
- **Informações dos coletores**: foram utilizadas as bases de dados de anúncios BGP de 200 pontos que fornecem informações aos coletores [60,66] selecionados de acordo com o descrito em [2] para um período de 300 dias imediatamente anteriores ao dia de teste, resultando em 1.024.242 enlaces direcionais únicos durante todo o período analisado.
- **Conjunto de dados para treinamento**: são geradas 1.000 amostras para cada uma das duas classes (legítimos/suspeitos) por dia para um período de 60 dias consecutivos imediatamente anteriores ao dia de teste, totalizando 120 mil amostras;
- **Novos enlaces**: são aqueles enlaces identificados no dia de teste e que não foram observados no período de 300 dias anteriores ou que não foram adicionados a base para esse período por serem considerados suspeitos;
- **Amostras para avaliação dos modelos**: foram geradas 1.000 amostras para cada uma das classes para cada dia de teste para se calcular as métricas de desempenho precisão, *recall* e *F1-score*. Essas amostras são distintas das utilizadas para o treinamento.

6.4 Avaliação dos Modelos

Para escrutinizar o modelo de DFOH e definir novos modelos, a ferramenta Trustee foi executada com os seguintes parâmetros: $M = \{0,2; 0,3; 0,4; 0,5; 0,6; 0,7; 0,8\}$, $N = \{40, 50\}$, $S = \{5, 10\}$ e $k = \{10, \infty\}$, resultando em 56 árvores de decisão, sendo 28 com poda e 28 sem poda. A relevância de cada *feature* foi determinada pela frequência de sua presença nas árvores de decisão geradas por Trustee, sendo as mais frequentes consideradas mais relevantes. A Tabela 6.2 mostra a distribuição das *features* na árvores de decisão sem poda ($k = \infty$) e a Tabela 6.3 nas árvores com poda.

Com base nesse critério, foram definidos os seguintes modelos:

- **M1**: modelo formado com todas as *features* originalmente utilizadas por DFOH;

Tabela 6.2: *Features* de DFOH utilizadas nas árvores de decisão sem poda.

Ordem	Feature	Árvores	Total	Topo
1	cone_degree	28	679	28
2	country_dist	28	845	0
3	degree	28	520	0
4	clustering_as1	28	483	0
5	ixp_dist	28	476	0
6	clustering_as2	28	430	0
7	average_neighbor_degree_as2	28	370	0
8	preferential_attachement	28	366	0
9	cone	28	365	0
10	jaccard	28	321	0
11	nb_vps	28	307	0
12	average_neighbor_degree_as1	28	302	0
13	closeness_centrality_as2	28	264	0
14	harmonic_centrality_as2	28	243	0
15	facility_country_dist	28	231	0
16	closeness_centrality_as1	28	229	0
17	facility_fac_dist	28	224	0
18	facility_cities_dist	28	212	0
19	triangles_as1	28	173	0
20	bidi	28	90	0
21	triangles_as2	27	147	0
22	harmonic_centrality_as1	24	145	0
23	eccentricity_as2	22	45	0
24	shortest_path	22	34	0
25	eccentricity_as1	17	22	0
26	degree_centrality_as1	0	0	0
27	degree_centrality_as2	0	0	0
28	adamic_adar	0	0	0

- **M2**: modelo original excluindo-se as *features* *degree_centrality_as1*, *degree_centrality_as2* e *adamic_adar*, que não apareceram em nenhuma das 28 árvores de decisão sem poda;
- **M3**: modelo M2 menos as três *features* seguintes que menos apareceram nas árvores sem poda, *eccentricity_as1*, *eccentricity_as2* e *shortest_path*;
- **M4**: modelo original de DFOH excluindo as 17 *features* que apareceram em no máximo uma das árvores com poda, permanecendo apenas as 11 *features* seguintes: *cone_degree*, *country_dist*, *clustering_as1*, *clustering_as2*, *bidi*, *triangles_as1*, *triangles_as2*, *preferential_attachement*, *degree*, *nb_vps* e *ixp_dist*; e
- **M5**: modelo formado com as cinco *features* que mais apareceram nas árvores com poda: *cone_degree*, *country_dist*, *clustering_as1*, *clustering_as2* e *bidi*.

Os cinco modelos foram inicialmente avaliados para as métricas de precisão, *recall* e *F1-Score* utilizando 10 conjuntos de teste, cada um com mil enlaces legítimos e mil forjados. O cálculo das *features* foi realizado para o dia primeiro de dezembro de 2022, usando os conjuntos de dados descritos na Seção 6.3. A Tabela 6.4 mostra o valor médio dos resultados para os 10 conjuntos de teste com o intervalo de confiança (IC) de 95%. Pode-se observar que não há diferença estatisticamente significativa entre os modelos. Assim, foram selecionados os modelos M4 e M5, por possuírem menos *features*, para uma comparação mais detalhada com M1.

Tabela 6.3: *Features* de DFOH utilizadas nas árvores de decisão com poda.

Ordem	Feature	Árvores	Total	Topo
1	cone_degree	28	60	28
2	country_dist	28	64	0
3	clustering_as1	27	39	0
4	clustering_as2	19	22	0
5	bidi	11	11	0
6	triangles_as1	6	6	0
7	preferential_attachement	4	4	0
8	ixp_dist	3	4	0
9	degree	3	3	0
10	nb_vps	3	3	0
11	triangles_as2	3	3	0
12	average_neighbor_degree_as2	1	1	0
13	harmonic_centrality_as2	1	1	0
14	cone	0	0	0
15	facility_fac_dist	0	0	0
16	facility_country_dist	0	0	0
17	facility_cities_dist	0	0	0
18	degree_centrality_as1	0	0	0
19	degree_centrality_as2	0	0	0
20	average_neighbor_degree_as1	0	0	0
21	eccentricity_as1	0	0	0
22	eccentricity_as2	0	0	0
23	harmonic_centrality_as1	0	0	0
24	closeness_centrality_as1	0	0	0
25	closeness_centrality_as2	0	0	0
26	shortest_path	0	0	0
27	jaccard	0	0	0
28	adamic_adar	0	0	0

Tabela 6.4: Médias de dez testes dos modelos, com os respectivos intervalos de confiança (IC) de 95%, para as métricas de precisão, *recall* e *F1-score*.

	Legítimos						**Suspeitos**					
M	**Precisão ± IC**		**Recall ± IC**		**F1-Score ± IC**		**Precisão ± IC**		**Recall ± IC**		**F1-Score ± IC**	
M1	0,9059	0,0054	0,9831	0,0029	0,9429	0,0035	0,9815	0,0032	0,8978	0,0064	0,9378	0,0041
M2	0,9059	0,0054	0,9830	0,0029	0,9429	0,0035	0,9814	0,0031	0,8979	0,0064	0,9378	0,0041
M3	0,9062	0,0053	0,9831	0,0028	0,9431	0,0034	0,9816	0,0031	0,8982	0,0063	0,9380	0,0039
M4	0,9062	0,0051	0,9829	0,0029	0,9430	0,0034	0,9814	0,0031	0,8982	0,0061	0,9379	0,0040
M5	0,9060	0,0052	0,9831	0,0028	0,9430	0,0033	0,9815	0,0030	0,8979	0,0062	0,9378	0,0039

A Tabela 6.5 mostra que houve redução de *features* em todas as categorias. Porém, não houve eliminação de nenhuma categoria, o que mostra que todas as categorias contribuem para o processo de decisão.

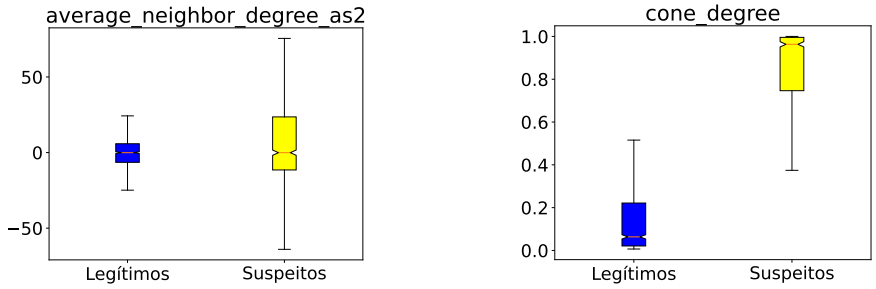
6.4.1 Impacto das *Features*

A ausência de diferença estatisticamente significativa entre os modelos mesmo com a remoção de várias *features* pode ser justificada por pelo menos dois motivos. Primeiro, pela *feature cone_degree*, que aparece na raiz da árvore de decisão da Figura 6.1. Ao calcular as métricas de desempenho de um modelo com apenas essa *feature* utilizando os valores presentes na árvore de decisão, verifica-se que ela sozinha atinge valores de precisão e *recall* acima de 0,9 para ambas as classes. Das 4.319 amostras legítimas, ela identifica corretamente 4.044, e das 4.499 amostras suspeitas, ela identifica corretamente 4.094.

Tabela 6.5: Comparativo da quantidade de *features* entre os modelos M1, M4 e M5.

Categoria	M1	M4	Redução M4	M5	Redução M5
Topológicas	18	5	72,22%	2	88,89%
Peering	5	2	60,00%	1	80,00%
Padrão do AS-path	3	2	33,33%	1	66,67%
Bidirecionalidade	2	2	0,00%	1	50,00%
Total	28	11	60,71%	5	82,14%

Segundo, algumas *features* impactam no desempenho dos modelos de forma mais significativa do que outras. Para se entender o motivo de algumas *features* terem grande relevância para o modelo e outras não, foram calculadas as distribuições dos valores dessas *features*. A Figura 6.2a mostra a distribuição dos valores de uma *feature* que não foi utilizada nos modelos reduzidos. Pode-se observar que os valores para a classe de enlaces legítimos estão contidos nos valores da classe de enlaces suspeitos, tornando a *feature* irrelevante para a separação das classes. Por outro lado, a Figura 6.2b mostra a distribuição de uma *feature* que foi incluída nos modelos reduzidos. Nota-se que os valores para as duas classes são bem distintos, contribuindo, portanto, para a separação das classes. Situação semelhante foi observada para as demais *features* que foram removidas ou mantidas nos modelos reduzidos.



(a) (b)

Figura 6.2: Boxplots com as distribuições dos valores de duas *features* entre as classes para as amostras geradas. A *feature* de (a) faz parte somente do modelo M1 enquanto a *feature* de (b) faz parte de todos os modelos.

Avaliação Qualitativa das *Features* Removidas de M4 e M5

As *features* para os modelos M4 e M5 foram escolhidas com base no número de vezes que elas apareceram nas 56 árvores de decisão geradas por Trustee. Entretanto, esse critério quantitativo não explica a razão das *features* removidas não contribuírem significativamente no processo de decisão dos modelos. A seguir é apresentada uma analise qualitativa que auxilia no entendimento da falta de contribuição de algumas *features*.

- *Padrão AS-path*: essa categoria é composta por três *features*, sendo que uma delas, *cone_ degree*, é gerada por uma floresta aleatória que possui como entrada os mesmos

valores utilizados na geração das outras duas. Assim, ela contém informações das outras duas em seu valor.

- *Topológicas*: essa categoria é dividida em subcategorias, como mostrado na Tabela 6.1. As *features* relacionadas a Centralidade, Distância (Dist.), Proximidade (Prox.) e Vizinhos se baseiam na quantidade de ASes conectados a um determinado AS e no comprimento dos *AS-paths*. Com o aumento dos pontos de troca de tráfego (PTTs) e de ASes conectados a eles, a hierarquia da Internet está se tornando mais plana, reduzindo portanto a relevância dessas *features*. Os ASes *stubs* se destacam em relação aos demais tipos de AS por terem uma única ligação, mas essa diferença também é capturada pelas *features* de Padrão Topológico (P. Topo.), que se mostraram mais relevantes. As *features clustering_as1* e *clustering_as2* se mostraram mais relevantes que as *triangles_as1* e *triangles_as2*, provavelmente por usar o valor das últimas duas em seus cálculos.

- *Peering*: As informações utilizadas para cálculo das *features* dessa categoria são fornecidas pelos administradores dos ASes e podem estar incompletas e desatualizadas, não sendo totalmente confiáveis para caracterização do AS [28]. A informação mais perene e confiável é o país onde o AS possui registro e está na *feature country_dist* que foi mantida em todos os modelos.

- *Bidirecionalidade*: uma ligação observada de forma bidirecional é suficiente para ser caracterizada como legítima [39, 72], tornando esta *feature* altamente relevante para as inferências. A *feature nb_vps*, que é calculada no código juntamente com a *feature bidi*, representa ligações com um grupo de ASes e pode estar perdendo a relevância com o aumento dos PTTs.

6.4.2 Tempo de Execução e Espaço Utilizado

DFOH é um sistema bastante modular e possui um módulo separado para calcular as *features* de cada categoria. Além disso, ele contém um módulo *broker* que calcula as *features* dos novos enlaces observados por dia, realiza o treinamento do modelo e a inferência das classes dos novos enlaces com o modelo treinado. A Tabela 6.6 mostra os tempos médios de execução, com os respectivos intervalos de confiança de 95%, dos módulos de DFOH para o período de 40 dias analisados. Pode-se observar que há uma redução em relação ao tempo de processamento de M1 para o cálculo das *features* superior a 31% (9 min e 44 seg) e 37% (11 min e 35 seg) para os modelos M4 e M5 (linha "Soma parcial" da tabela) e superior a 32% (5min e 46 seg) e 38% (6 min e 51 seg) para a execução do módulo *broker* (linha "Broker").

Os experimentos desta seção geraram arquivos de *features* para 160 dias (60 dias para o treinamento e 20 dias de testes para os dois períodos analisados) separados por categoria e tipos de amostra. Para as *features* de padrão de *AS-path* ainda é gerado um modelo por dia para cada uma das *features*, sendo três *features* para M1, duas para M4 e uma para M5. A Tabela 6.7 mostra o espaço de armazenamento necessário para processamento

Tabela 6.6: Comparação entre os tempos médios para execução dos modelos M1, M4 e M5. Os valores do IC estão em segundos e os valores em porcentagem correspondem a redução de tempo gasto em comparação com M1 assim como a diferença (Dif.) de tempo.

Módulo	M1		M4				M5			
	Tempo	IC	Tempo	IC	Dif. %	Dif. (M1)	Tempo	IC	Dif. %	Dif.(M1)
Padrão *AS-path*	2m e 48s	1	2m e 4s	1	26,32%	0m e 44s	1m e 37s	1	41,96%	1m e 11s
Bidirecionalidade	0m e 51s	8	0m e 50s	7	0,34%	0m e 1 s	0m e 46s	7	9,10%	0m e 5s
Peering	3m e 49s	5	0m e 56s	1	75,59%	2m e 53s	0m e 37s	1	83,87%	3m e 12s
Topológicas	23m e 47s	47	17m e 41s	41	25,67%	6m e 6s	16m e 40s	37	29,93%	7m e 7s
Soma Parcial	31m e 15s	45	21m e 31s	36	31,15%	9m e 44s	19m e 40s	32	37,04%	11m e 35s
Broker	17m e 44s	84	11m e 58s	64	32,55%	5m e 46s	10m e 53s	61	38,60%	6m e 51s

do modelo. Comparando o volume de dados utilizado para armazenamento dos arquivos relativos às *features* necessárias para os treinamentos dos modelos (diretório *features*), houve um redução de quase 60% no uso do espaço de armazenamento para M4 (82,3 MB) e de 71% (97,8 MB) para o M5 em relação ao espaço ocupado pelo modelo M1. Considerando o espaço necessário para as *features* e o espaço para os modelos, a redução fica em 36% (970,1 MB) para M1 e em 68% (1823,7 MB) para M5.

Tabela 6.7: Comparação entre o espaço de armazenamento necessário para armazenar as *features* dos modelos M1, M4 e M5. Os valores estão em bytes.

Diretório do sistema	M1	M4	Redução M4	M5	Redução M5
aspath_models_clusters	2651986704	1721000442	35,11%	842243747	68,24%
features	144141619	57872896	59,85%	41575973	71,16%
features/negative/aspath	11616412	8401632	27,67%	5193468	55,29%
features/negative/bidirectionality	2708608	2708609	0,00%	2269618	16,21%
features/negative/peeringdb	17637457	8240107	53,28%	5103018	71,07%
features/negative/topological	41857488	10378604	75,20%	8646301	79,34%
features/positive/aspath_clusters	11131982	8070168	27,50%	5046440	54,67%
features/positive/bidirectionality_clusters	2714919	2714914	0,00%	2316226	14,69%
features/positive/peeringdb_clusters	16351922	7594905	53,55%	4784388	70,74%
features/positive/topological_clusters	40106447	9747573	75,70%	8196034	79,56%

6.4.3 Avaliação das Métricas de Desempenho dos Modelos

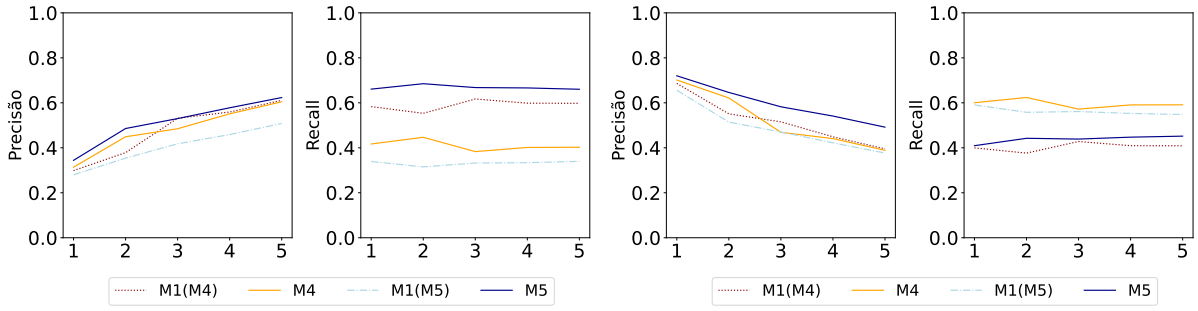
Apesar dos ganhos em tempo de processamento e espaço de armazenamento apresentados na seção anterior, um modelo com um conjunto menor de *features* deve ter desempenho semelhante ao original para ser útil. Os modelos foram analisados com enlaces forjados sintéticos, usando o mesmo procedimento descrito em [39]. Foram geradas mil amostras de enlaces legítimos e mil de suspeitos para cada um dos 40 dias avaliados. As *features* de cada um dos modelos M1, M4 e M5 foram calculadas para cada um dos conjuntos de amostra. Os valores médios das métricas de desempenho estão na Tabela 6.8 e mostram que os valores obtidos para M1 e M4 ficaram praticamente iguais e estão nos intervalos de confiança um do outro. Entretanto, os valores obtidos por M5 ficaram um pouco abaixo dos valores de M1, com diferenças de 0,0092 e 0,0091 nas médias do F1-score dos enlaces legítimos e suspeitos, respectivamente, e com distância máxima entre os intervalos de confiança de 0,0063.

Durante o período de análise, foram observados 16.107 novos enlaces únicos, dos quais 968 tiveram inferências diferentes entre M1 e M4, e 1.256 entre M1 e M5,

Tabela 6.8: Métricas para os modelos M1, M4 e M5, calculadas com base nas amostras geradas para os 40 dias de análise. São apresentados os valores médios, com os respectivos intervalos de confiança (IC) de 95%.

	Legítimos						**Suspeitos**					
M	**Precisão ± IC**		**Recall ± IC**		**F1-Score ± IC**		**Precisão ± IC**		**Recall ± IC**		**F1-Score ± IC**	
M1	0,9570	0,0018	0,9590	0,0022	0,9580	0,0014	0,9590	0,0021	0,9569	0,0019	0,9579	0,0013
M4	0,9582	0,0020	0,9580	0,0022	0,9581	0,0015	0,9581	0,0021	0,9581	0,0021	0,9581	0,0015
M5	0,9488	0,0020	0,9488	0,0025	0,9488	0,0015	0,9488	0,0024	0,9488	0,0021	0,9488	0,0015

correspondendo a 6,0% e 7,8%, respectivamente. Como não há uma verdade absoluta (*ground truth*) para determinar qual modelo é melhor, considerou-se que sequestros de prefixo geralmente ocorrem por curtos períodos, conforme relatado por [81] e observado em incidentes noticiados [70, 78]. Assim, foram analisadas as RIBs das primeiras duas horas de janeiro a maio de 2023 e 2024 para verificar se esses enlaces com divergências nas classificações apareceram nesses períodos, ou seja, posteriormente ao suposto sequestro. Os enlaces que apareceram foram considerados legítimos e os demais como suspeitos. A Figura 6.3 mostra os resultados para os enlaces com divergências nas classificações entre os modelos considerando o número de dias que eles apareceram nas RIBs para que fossem classificados como legítimos. M1(M4) e M1(M5) representam os resultados do modelo M1 para os enlaces que tiveram inferências divergentes entre o modelo M1 e os modelos M4 e M5, respectivamente.



(a) Suspeitos (b) Legítimos

Figura 6.3: Métricas calculadas com base no número mínimo de dias (RIBs) em que um enlace é observado para ser classificado como legítimo. M1(M4) e M1(M5) representam os resultados do modelo M1 para enlaces com inferências divergentes em relação aos modelos M4 e M5, respectivamente.

Um sistema de identificação de sequestros deve minimizar falsos positivos para manter sua credibilidade junto aos operadores de rede [39]. Nesse contexto, o modelo M5 apresentou resultados melhores que o modelo M1. Independentemente do número de dias que um enlace é observado para considerá-lo legítimo, o modelo M5 foi mais preciso em ambas as classes, sendo 0,1152 mais preciso que o modelo M1 ao considerar enlaces observados por 5 dias em meses consecutivos. Como algumas *features* são baseadas em informações fornecidas pelos operadores e outras podem estar perdendo a relevância devido às mudanças na topologia da Internet que a tem tornado mais "plana" (*i.e.*, caminhos cada vez mais curtos entre os ASes), elas podem estar contribuindo

negativamente para o modelo original, sendo uma das prováveis justificativas para a melhora de desempenho dos modelos reduzidos e reforçando a necessidade de análises mais detalhadas na seleção de *features*.

6.5 Análise das *Features* com XAI Local

O uso de técnicas de explicabilidade local permite analisar a relevância das *features* para cada inferência, determinando sua influência em cada classe e o grau de interferência. Nesta seção, aplicou-se a ferramenta Lime [68] às inferências de todas as amostras usadas na avaliação dos modelos M1, M4 e M5 ao longo de 40 dias, com mil amostras por classe por dia. A Tabela 6.5 apresenta as *features* ordenadas pela média da contribuição de cada uma em todas as inferências realizadas, com Intervalo de Confiança (IC) de 95

Tabela 6.9: *Features* ordenadas com base na média do valor que colaborou para cada inferência de uma forma geral, para a classe suspeito (sus) ou para classe legítimo (leg). Valores calculados para as amostras utilizadas para validação dos 40 dias de teste.

Ord	Feature	M4	M5	Média	IC	Média (sus)	IC (sus)	Média (leg)	IC (leg)
1	bidi	X	X	0,4647	0,0002	0,4649	0,0002	-0,4557	0,0017
2	cone_degree	X	X	0,4176	0,0006	0,4177	0,0010	-0,4176	0,0007
3	country_dist	X	X	0,1955	0,0007	0,1952	0,0012	-0,1958	0,0005
4	degree	X		0,0832	0,0002	0,0832	0,0004	-0,0832	0,0003
5	clustering_as2	X	X	0,0719	0,0003	0,0558	0,0003	-0,1031	0,0006
6	cone			0,0611	0,0002	0,0625	0,0004	-0,0598	0,0002
7	clustering_as1	X	X	0,0547	0,0003	0,0411	0,0002	-0,0848	0,0006
8	ixp_dist	X		0,0430	0,0002	0,0663	0,0004	-0,0317	0,0001
9	nb_vps	X		0,0276	0,0001	0,0276	0,0002	-0,0276	0,0002
10	facility_country_dist			0,0262	0,0001	0,0299	0,0002	-0,0233	0,0001
11	average_neighbor_degree_as2			0,0210	0,0001	0,0204	0,0001	-0,0217	0,0002
12	jaccard			0,0204	0,0001	0,0176	0,0001	-0,0245	0,0002
13	triangles_as2	X		0,0181	0,0001	0,0197	0,0001	-0,0165	0,0001
14	triangles_as1	X		0,0178	0,0001	0,0193	0,0001	-0,0163	0,0001
15	closeness_centrality_as1			0,0164	0,0001	0,0161	0,0001	-0,0168	0,0001
16	average_neighbor_degree_as1			0,0155	0,0001	0,0152	0,0001	-0,0159	0,0001
17	facility_cities_dist			0,0149	0,0001	0,0149	0,0001	-0,0149	0,0001
18	closeness_centrality_as2			0,0141	0,0001	0,0139	0,0001	-0,0144	0,0001
19	harmonic_centrality_as1			0,0131	0,0001	0,0131	0,0001	-0,0132	0,0001
20	facility_fac_dist			0,0130	0,0001	0,0131	0,0001	-0,0130	0,0001
21	preferential_attachement	X		0,0130	0,0001	0,0129	0,0001	-0,0132	0,0001
22	harmonic_centrality_as2			0,0129	0,0001	0,0129	0,0001	-0,0130	0,0001
23	degree_centrality_as1			0,0114	0,0001	0,0114	0,0001	-0,0114	0,0001
24	degree_centrality_as2			0,0114	0,0001	0,0114	0,0001	-0,0114	0,0001
25	adamic_adar			0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
26	eccentricity_as1			0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
27	eccentricity_as2			0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
28	shortest_path			0,0000	0,0000	0,0000	0,0000	0,0000	0,0000

Na Tabela 6.5, nota-se que as cinco *features* do modelo M5 estão entre as sete mais relevantes, enquanto as *features* adicionais de M4 aparecem mais dispersas. As quatro *features* que não contribuíram para as inferências também não apareceram nas árvores podadas geradas pela ferramenta Trustee, posicionando-se entre as seis menos relevantes nas árvores completas. Segundo Lime, as *features* `bidi` e `cone_degree` apresentaram relevância significativamente superior. A `cone_degree` já havia mostrado importância em

análises anteriores, enquanto a relevância de `bidi` foi destacada em [39, 72] e confirmada nesta análise.

A Figura 6.4 ilustra visualmente os resultados do Lime, evidenciando que o peso de cada *feature* varia entre as inferências. As barras vermelhas representam a contribuição para a classe de enlaces legítimos (negativos) e as verdes, para a classe de suspeitos (positivos).

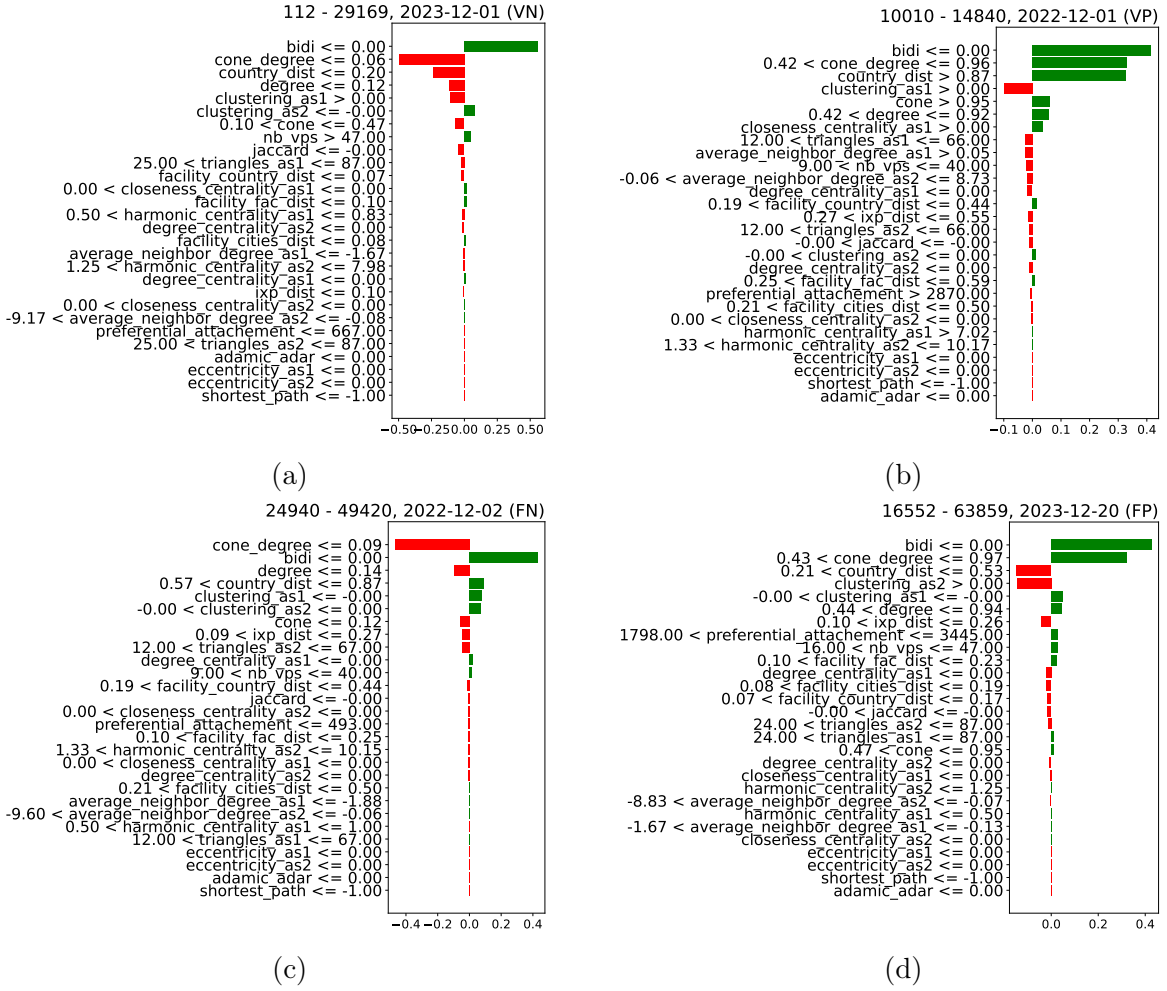


Figura 6.4: Exemplo de análise gerada pelo Lime. O título de cada imagem corresponde ao enlace observado (AS1-AS2), a data em que o enlace foi observado (ano-mês-dia) e o resultado da inferência, sendo em (a) Verdadeiro Negativo (VN), em (b) Verdadeira Positiva (VP), em (c) Falso Negativa (FN) e em (d) Falso Positiva (FP).

Lime foi usada para comparar a relevância das *features* com as identificadas pela ferramenta Trustee, sem diferenças que invalidassem a análise inicial, corroborando as *features* mais relevantes. Como trabalho futuro, os resultados de Lime poderão ser usados para analisar e corrigir inferências incorretas.

6.6 Identificação dos Sequestros da Caracterização

Para verificar a capacidade de detecção dos sequestros em ASes militares pelo DFOH, os *AS-paths* sequestrados nas simulações observados pelos coletores foram convertidos para o formato do DFOH, simulando novos enlaces. Foram realizados 17.400 sequestros de Tipo-1 para cada data analisada. No entanto, como destacado na Seção 5.4, uma porcentagem considerável de sequestros não chega aos coletores, o que impossibilita análises desses sequestros com DFOH e que não seriam, portanto, identificados. Para os enlaces observados pelos coletores, o modelo M1 identificou ∼94% dos sequestros, o M4 ∼93% e o modelo M5 ∼90%. Entretanto, se for considerado todos os sequestros simulados, a taxa de identificação cai ∼5% devido aos sequestros não observados. A Tabela 6.10 detalha os valores.

Tabela 6.10: Número de inferências corretas, verdadeiros positivos (VP), realizadas pelo DFOH e o valor porcentual em relação aos enlaces observados pelos coletores (Obs.) e em relação ao total de simulações para os modelos M1, M4 e M5 (17.400 no total).

		M1			**M4**			**M5**		
Data	**Obs.**	**VP**	**Obs.**	**Total**	**VP**	**Obs.**	**Total**	**VP**	**Obs.**	**Total**
01-02-2024	16.331	15.399	94,41%	88,50%	15.231	93,38%	87,53%	14.727	90,29%	84,64%
01-03-2024	16.428	15.502	94,36%	89,09%	15.301	93,14%	87,94%	14.772	89.92%	84,90%
01-04-2024	16.474	15.497	94,07%	89,06%	15.363	93,26%	88,29%	14.929	90,62%	85,80%

A vantagem de utilização dos modelos reduzidos para inferência dos dados obtidos na simulação ficou no quesito de tempo de execução. Para a simulação com data de primeiro de abril de 2024, o tempo necessário para o cálculo das *features* de M1 foi de 19.182 segundos (5 horas 19 minutos e 42 segundos), M4 14.294 segundos (3 horas 58 minutos e 14 segundos) e M5 12.760 segundos (3 horas 32 minutos e 40 segundos). A redução do tempo necessário ficou em 25,48% (1 hora 21 minutos e 28 segundos) e 33,48% (1 hora 47 minutos e 2 segundos), comparando os modelos M4 e M5 ao M1, respectivamente.

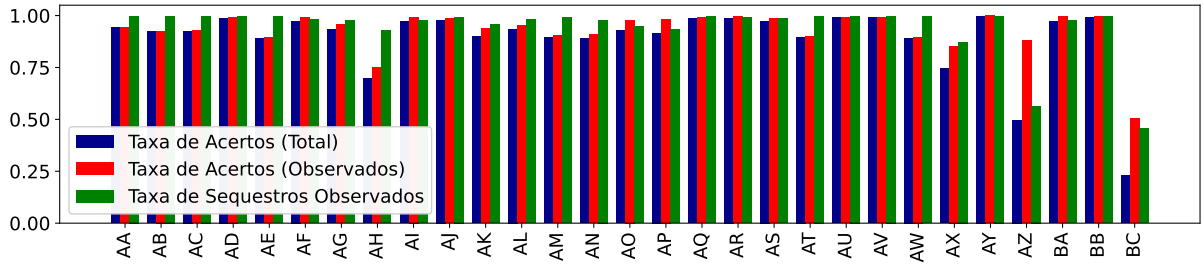


Figura 6.5: Taxa de acerto da inferência do DFOH por AS para a simulação com base no dia 01-04-2024 em relação a quantidade de sequestros total e observados pelos coletores. A taxa de sequestros observados pelos coletores para casa AS também é apresentada

A Figura 6.5 apresenta as taxas de acerto do modelo M1 com base no total de sequestros simulados por AS e nos sequestros observados pelos coletores, além da taxa de enlaces observados. Muito embora a Seção 5.5 mostre que os ASes AZ e BC sejam os

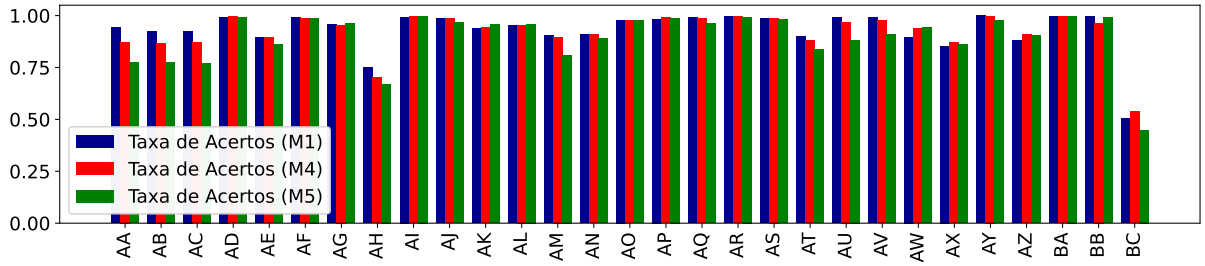


Figura 6.6: Taxa de acerto em relação aos enlaces observados por modelo, M1, M4 e M5

mais resilientes, eles também são mais suscetíveis a sequestros invisíveis, ou seja, sequestros não observados nos coletores ou detectados pelo DFOH.

O gráfico da Figura 6.6 apresenta a taxa de acerto para os enlaces observados por modelo. Pode-se observar que o modelo M1, que apresentou um resultado geral melhor, não foi o que mais detectou os sequestros para todos os ASes vítimas. O modelo M1 apresentou melhor resultado para 17 ASes, sendo que para 4 teve o mesmo resultado que M4. O modelo M4 acertou mais para 10 ASes, contando os 4 em que empatou com M1, para outros 4 ASes ele teve desempenho melhor que M1, mas inferior ao M5. Já o modelo M5 teve melhor resultado para 6 ASes, sendo que em outros 3 ele ficou melhor que M1, mas não que o M4. Considerando-se a diferença de tempo de execução que M4 gastou a menos que M1 para o cálculo das *features* (1 hora 21 minutos e 28 segundos) e que dos 29 ASes vítimas o modelo M4 teve resultado melhor ou igual a M1 para 14 ASes (48,3%), mostra que seu uso pode ser mais viável em algumas situações onde há muitos enlaces a ser inferido e há a necessidade de uma resposta mais rápida.

Como conclusão deste estudo, mais de 10% de sequestros Tipo-1 não seriam detectados por nenhum dos modelos apresentados para a melhor ferramenta existente atualmente, sendo que alguns ASes poderiam ter mais de 77% de sequestros não identificados ou visíveis nos coletores públicos de rota.

6.7 Melhorias Propostas para o Modelo

A seguir são apresentadas duas propostas para aprimorar o modelo utilizado por DFOH: a primeira explora o aumento de fontes de dados para a geração das *features*, enquanto a segunda propõe alterações nas amostras utilizadas no treinamento.

6.7.1 Validação de Novas Fontes de Dados

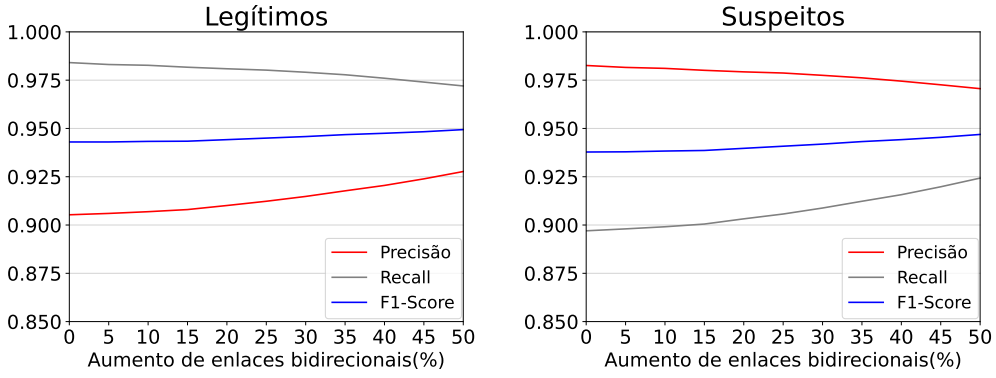
Para melhorar o modelo, foi avaliado o impacto do incremento de informações na geração da *feature* `bidi`, identificada como relevante em estudos anteriores. A *feature* `bidi` indica enlaces observados nos dois sentidos nos *AS-path* analisados, que, segundo pesquisas [39, 72], não podem ser forjados por serem bidirecionais. Dados do plano de dados podem ser usados para identificar enlaces bidirecionais. Atualmente, apenas uma pequena fração dos enlaces monitorados por DFOH é bidirecional (7%): 18.310 de 356.865 em primeiro

de dezembro de 2022, e 38.981 de 620.397 em primeiro de dezembro de 2023.

Para avaliar o impacto do aumento dessa *feature* na inferência dos enlaces, foram realizados os seguintes experimentos:

1. Geração de todas as *features* necessárias para o treinamento dos modelos tendo com base o dia primeiro de dezembro de 2022;
2. Treinamento do modelo com o código original e sem alteração das *features*;
3. Incremento na quantidade de enlaces bidirecionais para as ligações legítimas entre os ASes, de forma aleatória, no valor de 5 à 50%, aumentando em 5% a cada execução;
4. Treinamento de novos modelos para cada valor de incremento;
5. Os passos 3 e 4 foram executados 10 vezes, de forma a se obter 10 modelos treinados com incrementos diferentes para cada porcentagem de aumento;
6. Foram gerados mil enlaces de cada classe para validação dos modelos, análogo à validação original da ferramenta DFOH;
7. A seleção de enlaces gerados foi inferida por todos os modelos gerados.

Esses passos permitem analisar como o aumento da *feature* `bidi` afeta a precisão e o desempenho do modelo.



(a) (b)

Figura 6.7: Análise da precisão, *recall* e F1-*score* com o aumento dos enlaces bidirecionais. Em (a) o resultado para classe Legítimos e em (b) para classe Suspeitos.

Após as inferências realizadas, foram calculados os valores de precisão, *recall* e *F1-score*. A Figura 6.7a apresenta a média dos resultados obtidos em 10 inferências realizadas por 10 modelos distintos para cada porcentagem de incremento nos enlaces inferidos como legítimos. O valor correspondente ao modelo original, sem incremento, está representado no ponto 0 do eixo X. Já a Figura 6.7b apresenta os valores para os enlaces inferidos como suspeitos.

Na Figura 6.7a, observa-se um aumento na precisão para enlaces legítimos acompanhado de uma queda no *recall*, enquanto na Figura 6.7b ocorre o oposto: um

aumento no *recall* e redução na precisão para enlaces suspeitos. Em ambas as classes, o *F1-score* apresenta um leve aumento. Esses resultados indicam que o incremento aleatório na *feature* `bidi` teve impacto positivo na inferência dos enlaces.

Uma nova fonte de dados capaz de aumentar a quantidade de enlaces bidirecionais observados poderia melhorar a inferência do modelo, tanto ao incrementar os enlaces bidirecionais usados no treinamento, como nesta análise, quanto ao identificar mais enlaces bidirecionais durante as inferências. Essa característica, como já mencionado, é um indicativo de enlaces legítimos. Fontes adicionais de dados podem ser obtidas por meio de projetos que mapeiam caminhos na Internet no plano de dados, utilizando ferramentas como o `traceroute`. Contudo, essa análise será foco de trabalho futuro.

6.7.2 Validação de Novas Amostras de Treinamento

A Tabela 6.8 mostra que os modelos analisados apresentaram métricas muito boas ao serem avaliados com enlaces forjados sinteticamente pelo sistema DFOH. No entanto, conforme observado na Figura 6.3, que considera apenas os enlaces com inferências divergentes entre os modelos, as métricas obtidas são inferiores quando analisados novos enlaces reais observados na Internet.

Com base na premissa de que os sequestros de prefixos ocorrem por períodos curtos, foram gerados novos conjuntos de treinamento para os modelos utilizando enlaces observados seis e sete meses antes do período de treinamento. Os enlaces suspeitos foram definidos como aqueles que não foram mais observados nos seis meses seguintes ao mês em que apareceram como novos. Para os enlaces considerados legítimos, foi aplicada uma variação na quantidade mínima de meses (de um a seis) em que o enlace precisou ser observado novamente para ser classificado como legítimo. As amostras foram balanceadas selecionando-se aleatoriamente enlaces de maior frequência para igualar as quantidades entre as classes. Os modelos resultantes foram denominados **ne**`X`, onde `X` indica a quantidade mínima de meses (de 1 a 6) em que o enlace foi observado para ser considerado legítimo.

A análise e validação desses novos modelos seguiram o mesmo período e metodologia descritos na Seção 6.3. A classificação dos enlaces como legítimos ou suspeitos foi realizada conforme a abordagem da Seção 6.4.3. Devido à maior quantidade de enlaces legítimos em comparação aos suspeitos, a métrica MCC foi utilizada para comparar os modelos. Os resultados estão apresentados na Tabela 6.11.

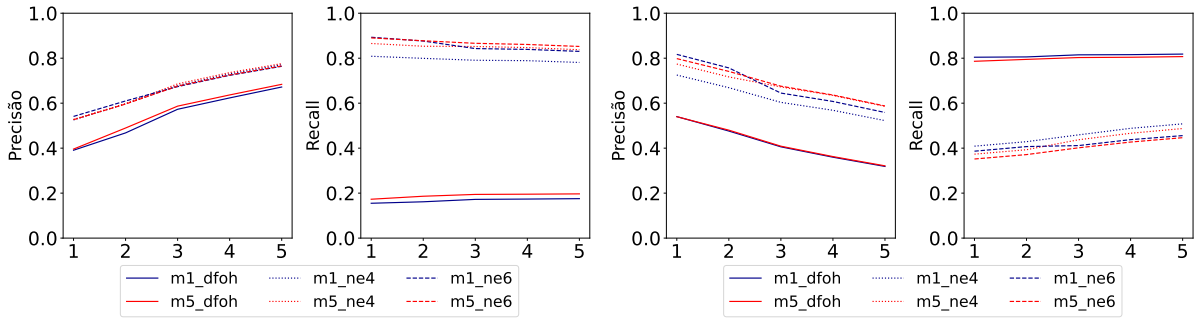
Os resultados da Tabela 6.11 mostram que os modelos treinados com a amostragem original (linha DFOH) apresentaram valores de MCC negativos ou próximos de zero, indicando desempenho equivalente ao de uma inferência aleatória, como o lançamento de uma moeda. Apesar disso, o modelo M1, treinado com o conjunto `ne6`, obteve valores de MCC significativamente superiores quando a validação de um novo enlace como legítimo requer sua observação por apenas um ou dois meses. Por outro lado, o modelo M5, com o conjunto `ne4`, apresentou melhores resultados quando exigida a observação de três a cinco meses para validar um enlace como legítimo.

Os valores de precisão e *recall* para os dois modelos com maior MCC estão apresentados na Figura 6.8, juntamente com os resultados do treinamento com a amostragem original

Tabela 6.11: Valores de MCC para os modelos M1, M4 e M5 para o treinamento com a amostragem original (DFOH) e para a amostragem com base nos novos enlaces (ne1 à ne6). Em destaque o maior valor para cada exigência de dias observados.

	Min 1 mês p/ leg			Min 2 meses p/ leg		
	M1	**M4**	**M5**	**M1**	**M4**	**M5**
DFOH	-0,0530	-0,0444	-0,0512	-0,0429	-0,0267	-0,0241
ne1	0,2159	0,1559	0,2339	0,2303	0,1794	0,2436
ne2	0,1314	0,1249	0,2416	0,1563	0,1512	0,2542
ne3	0,2287	0,0700	0,2392	0,2419	0,1013	0,2503
ne4	0,2335	0,1172	0,2685	0,2462	0,1415	0,2783
ne5	0,2872	0,2705	0,2690	0,2955	0,2776	0,2799
ne6	0,3165	0,2893	0,2803	0,3223	0,2960	0,2897

	Min 3 meses p/ leg			Min 4 meses p/ leg			Min 5 meses p/ leg		
	M1	**M4**	**M5**	**M1**	**M4**	**M5**	**M1**	**M4**	**M5**
DFOH	-0,0166	-0,0163	-0,0036	-0,0134	-0,0072	-0,0007	-0,0077	-0,0027	0,0044
ne1	0,2033	0,1613	0,2279	0,2317	0,1903	0,2540	0,2342	0,2020	0,2658
ne2	0,1547	0,1469	0,2825	0,1861	0,1774	0,3051	0,1973	0,1886	0,3116
ne3	0,2257	0,1546	0,2948	0,2527	0,1847	0,3149	0,2558	0,1953	0,3210
ne4	0,2654	0,1387	0,3219	0,2877	0,1687	0,3408	0,2913	0,1812	0,3429
ne5	0,2659	0,2906	0,3212	0,2858	0,3053	0,3380	0,2884	0,3046	0,3413
ne6	0,2846	0,2703	0,3067	0,3030	0,2876	0,3228	0,3038	0,2883	0,3249



(a) Suspeitos (b) Legítimos

Figura 6.8: Métricas calculadas com base no número mínimo de dias (RIBs) em que um enlace é observado para ser classificado como legítimo para os modelos M1 e M5 treinados com a amostragem original (dfoh) e com as amostras n4 e n6.

para comparação. Observa-se que o treinamento com novos enlaces apresenta melhor desempenho na precisão para ambas as classes e no *recall* da classe de suspeitos. O treinamento original destacou-se apenas no *recall* da classe de legítimos.

Devido à ausência de uma verdade absoluta para validar os enlaces, não é possível determinar qual modelo é efetivamente superior. No entanto, considerando a curta duração típica dos sequestros, como mencionado anteriormente, o uso de dados reais para o treinamento do modelo aparenta oferecer melhores resultados em comparação ao uso de enlaces forjados de forma sintética.

Capítulo 7

Conclusão

Sequestros de prefixo têm causado grandes problemas na Internet e, apesar das técnicas que visam aumentar a segurança do BGP, ainda devem ocorrer por muito tempo. No estudo de caracterização de sequestros de prefixo, pode-se verificar a gravidade dos sequestros de prefixo para ASes utilizados por Forças Armadas do G20. Esses ataques podem contaminar grande parte da Internet, permitindo ações como negação de serviço ou ataques *man-in-the-middle*, comprometendo a segurança e estabilidade dos ASes. Embora sequestros Tipo-0 tenham maior alcance, as simulações mostram que a implementação de RPKI reduz significativamente a vulnerabilidade dos ASes, sendo uma das formas mais eficazes de mitigação. No entanto, sua adoção ampla é essencial, pois a implementação parcial limita sua efetividade.

Sequestradores podem recorrer a *AS-paths* forjados para burlar RPKI, como observado em [78]. Dado que ainda não há uma técnica definitiva para prevenir esses ataques, o aumento da resiliência e a identificação precisa são estratégias fundamentais. Para melhorar a resiliência, este trabalho mostra que contratar provedores bem conectados e geograficamente distribuídos é crucial. O uso de *prepend*, especialmente na origem do anúncio, deve ser evitado, pois reduz a resiliência, principalmente para ASes com poucos provedores. Como medidas de mitigação, os ASes podem ampliar o número de pontos de divulgação de seus prefixos, aumentando a resiliência e reduzindo os impactos do ataque.

Muitos trabalhos buscam realizar a identificação do sequestro e modelos do tipo caixa-preta tem apresentado desempenhos melhores que modelos interpretáveis. No entanto, operadores de rede ainda são relutantes em utilizá-los em situações críticas. O uso de técnicas de XAI pode auxiliar a identificar a real relevância de cada *feature* e seu peso nas inferências, trazendo maior confiabilidade nas decisões do modelo. Além disso, elas podem mostrar que uma quantidade maior de *features* em um modelo não garante o seu sucesso, mas sim quão relevantes elas são. Selecionar o melhor conjunto de *features* pode economizar tempo de execução e espaço de armazenamento, além de poder até mesmo melhorar o desempenho de um modelo.

Com a avaliação experimental extensiva realizada neste trabalho, observou-se que um modelo com um conjunto de *features* 60,71% menor alcançou resultados sem diferenças estatisticamente significativas aos do modelo completo. Adicionalmente, um modelo com um conjunto de *features* 82,14% menor obteve métricas similares e maior

precisão com dados reais que o modelo original. Os modelos reduzidos também diminuíram o tempo de execução em mais de 30% e o espaço de armazenamento em mais de 59%. Por serem computacionalmente mais leves, eles podem ser retreinados com maior frequência e assim capturar mais rapidamente mudanças nas distribuições dos dados. Analisando os sequestros realizados na caracterização dos ASes militares, este trabalho mostra que mais de 10% dos sequestros não são detectados pela melhor ferramenta existente ou não são observados nos coletores públicos, sendo sua identificação essencial para mitigação. Uma das soluções para reduzir essa taxa seria o aumento de coletores em regiões estratégicas.

A avaliação experimental também mostra que, além da correta seleção *features*, mais informações para *features* relevantes, ou melhores amostras de treinamento, podem melhorar o modelo utilizado.

7.1 Trabalhos Futuros

Prosseguindo no estudo do sequestro de prefixos, as seguintes atividades são propostas para continuidade do trabalho:

- **Resiliência dos ASes**: Neste trabalho foi analisada a resiliência dos ASes militares. Novos estudos de caracterização podem ser realizados com outras classes de ASes, como ISP e/ou de serviços de hospedagem, como forma de auxiliar na contratação de serviços mais resilientes a sequestros.

- **Melhoria do modelo**: Conforme discutido na Seção 6.7.1, novas fontes de dados que possam caracterizar enlaces como sendo bidirecionais podem melhorar o desempenho do modelo. Fontes adicionais que possam enriquecer essa ou outras *features* podem produzir resultados melhores. Por exemplo, o uso de dados de `traceroute` disponibilizados em [15, 16, 37] pode revelar enlaces bidirecionais não observados nos coletores BGP;

- **Uso de XAI local**: Análises mais detalhadas das inferências incorretas com o uso de XAI local podem auxiliar na correção do modelo analisado e permitir a redução de falsos positivos;

Referências Bibliográficas

[1] M. Ahmed, R. Seraj, and S. M. S. Islam. The K-Means Algorithm: A Comprehensive Survey and Performance Evaluation. *Electronics*, 9(8):1295, 2020.

[2] T. Alfroy, T. Holterbach, and C. Pelsser. MVP: Measuring Internet Routing from the Most Valuable Points. In *Proceedings of the 22nd ACM Internet Measurement Conference*, IMC '22, page 770–771, 2022.

[3] I. A. N. Authority. About Us. `https://www.iana.org/about`. acessado em 08/11/2023.

[4] A. Azimov, E. Bogomazov, R. Bush, K. Patel, J. Snijders, and K. Sriram. BGP AS_PATH Verification Based on Autonomous System Provider Authorization (ASPA) Objects. Internet-Draft draft-ietf-sidrops-aspa-verification-20, Internet Engineering Task Force, Jan. 2025. Work in Progress.

[5] F. Baker. Requirements for IP Version 4 Routers. RFC 1812, June 1995.

[6] R. Beltiukov, W. Guo, A. Gupta, and W. Willinger. In Search of netUnicorn: A Data-Collection Platform to Develop Generalizable ML Models for Network Security Problems. In *Proc. of the 2023 ACM Conf. on Computer and Comm. Security*, CCS '23, page 2217–2231, 2023.

[7] H. Birge-Lee, Y. Sun, A. Edmundson, J. Rexford, and P. Mittal. Bamboozling Certificate Authorities with BGP. In *Proc. of the 27th USENIX Security'18*, pages 833–849, Aug. 2018.

[8] H. Birge-Lee, L. Wang, J. Rexford, and P. Mittal. SICO: Surgical Interception Attacks by Manipulating BGP Communities. In *Proc. of the 2019 ACM Conference on Computer and Communications Security*, CCS '19, page 431–448, 2019.

[9] H. Birge-Lee, J. Wanner, G. H. Cimaszewski, J. Kwon, L. Wang, F. Wirz, P. Mittal, A. Perrig, and Y. Sun. Creating a Secure Underlay for the Internet. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 2601–2618, 2022.

[10] T. Bühler, A. Milolidakis, R. Jacob, M. Chiesa, S. Vissicchio, and L. Vanbever. Oscilloscope: Detecting BGP Hijacks in the Data Plane. *arXiv preprint arXiv:2301.12843*, 2023.

[11] R. Bush and R. Austein. The Resource Public Key Infrastructure (RPKI) to Router Protocol, Version 1. RFC 8210, Sept. 2017.

[12] CAIDA. AS Relationships (Serial-1). `https://publicdata.caida.org/datasets/as-relationships/serial-1/`, 2013. acessado em 31/07/2024.

[13] CAIDA. AS Relationships (Serial-2). `https://catalog.caida.org/dataset/as_relationships_serial_2`, 2015.

[14] CAIDA. CAIDA AS Rank. `http://asrank.caida.org/`, 2022/2023. acessado em 31/07/2024.

[15] CAIDA. IPv4 Prefix-Probing Traceroute Dataset. `https://catalog.caida.org/dataset/ark_ipv4_prefix_probing`, 2025.

[16] CAIDA. IPv6 Prefix-Probing Traceroute Dataset. `https://catalog.caida.org/dataset/ark_ipv6_prefix_probing`, 2025.

[17] A. B. Carvalho, B. A. da Silva Jr, C. A. da Silva, and R. A. Ferreira. Material suplementar. https://github.com/Bastos-abc/blackbox-explainable-xai-hijack, 2024.

[18] A. B. Carvalho, P. B. Marcos, F. S. de Paula, C. A. da Silva, and R. A. Ferreira. Caracterização da Vulnerabilidade a Sequestros de Prefixo de Sistemas Autônomos Militares. *Submetido*, 2025.

[19] A. B. Carvalho, P. B. Marcos, F. S. de Paula, C. A. da Silva, and R. A. Ferreira. Hijack simulator. https://github.com/Bastos-abc/hijack_simulator, 2025.

[20] A. B. Carvalho, B. Silva Jr, C. Silva, and R. A. Ferreira. Abrindo a Caixa-Preta – Aplicando IA Explicável para Aprimorar a Detecçao de Sequestros de Prefixo. *Anais do XXIII Simpósio Brasileiro em Segurança da Informaçao e de Sistemas Computacionais. SBC*, 2024.

[21] Z. Chen, Z. S. Bischof, C. Testart, and A. Dainotti. Improving the Inference of Sibling Autonomous Systems. In A. Brunstrom, M. Flores, and M. Fiore, editors, *Passive and Active Measurement*, pages 345–372, Cham, 2023. Springer Nature Switzerland.

[22] D. Chicco and G. Jurman. The Advantages of the Matthews Correlation Coefficient (MCC) over F1 Score and Accuracy in Binary Classification Evaluation. *BMC genomics*, 21:1–13, 2020.

[23] S. Cho, R. Fontugne, K. Cho, A. Dainotti, and P. Gill. BGP Hijacking Classification. In *2019 Network Traffic Measurement and Analysis Conference (TMA)*, pages 25–32, 2019.

[24] I. Cisco Systems. Select BGP Best Path Algorithm. `https://www.cisco.com/c/en/us/support/docs/ip/border-gateway-protocol-bgp/13753-25.html`. acessado em 19/02/2024.

[25] Cybersecurity and Infrastructure Security Agency. Understanding and Responding to Distributed Denial-of-Service Attacks. `https://www.cisa.gov/sites/default/files/publications/understanding-and-responding-to-ddos-attacks_508c.pdf`, Oct. 2022. acessado em 20/02/2024.

[26] B. A. da Silva Jr, A. B. de Carvalho, I. Cunha, T. Friedman, E. Katz-Bassett, and R. A. Ferreira. Uncovering BGP Action Communities and Community Squatters in the Wild. *Proc. ACM Meas. Anal. Comput. Syst.*, 8(3), Dec. 2024.

[27] B. A. da Silva Jr, P. Mol, O. Fonseca, I. Cunha, R. A. Ferreira, and E. Katz-Bassett. Automatic Inference of BGP Location Communities. In *Proc. of ACM SIGMETRICS/IFIP Performance*, pages 3.1–3.23, Mumbai, India, June 2022.

[28] B. Du, K. Izhikevich, S. Rao, G. Akiwate, C. Testart, A. C. Snoeren, and k. claffy. IRRegularities in the Internet Routing Registry. In *Proc. of the ACM IMC 2023*, page 104–110, 2023.

[29] A. M. e Ricardo Patara. Alocação de Endereços IP e ASN para Provedores Internet. `https://www.nic.br/publicacoes/indice/guias/`, 2022. acessado em 08/11/2023.

[30] H. Electric. Hurricane Electric Internet Services. `https://bgp.he.net/`, 2024.

[31] O. Fonseca, I. Cunha, E. Fazzion, W. M. Jr., B. A. Silva Junior, R. A. Ferreira, and E. Katz-Bassett. Identifying Networks Vulnerable to IP Spoofing. *IEEE Transactions on Network and Service Management (TNSM)*, 18(3):3170–3183, September 2021.

[32] O. Fonseca, I. Cunha, E. Fazzion, W. M. Jr., B. A. Silva Junior, R. A. Ferreira, and E. Katz-Bassett. Tracking Down Sources of Spoofed IP Packets. In *Proceedings of IFIP Networking Conference (NETWORKING'20)*, pages 15–21, Paris, France, June 2029.

[33] D. Freedman, B. Foust, B. Greene, B. Maddison, A. Robachevsky, J. Snijders, and S. Steffann. Mutually Agreed Norms for Routing Security (MANRS) Implementation Guide, 2019.

[34] L. Gao and J. Rexford. Stable Internet Routing Without Global Coordination. *IEEE/ACM Transactions on Networking*, 9(6):681–692, 2001.

[35] A. Gavrichenkov. Breaking HTTPS with BGP hijacking. `https://www.blackhat.com/docs/us-15/materials/us-15-Gavrichenkov-Breaking-HTTPS-With-BGP-Hijacking-wp.pdf`. acessado em 20/02/2024.

[36] P. Geurts, D. Ernst, and L. Wehenkel. Extremely Randomized Trees. *Machine learning*, 63:3–42, 2006.

[37] D. group at Sorbonne Université. Iris. `https://iris.dioptra.io/`, 2025.

[38] B. Herdes, M. Zhang, and T. Ryan. Cloudflare 1.1.1.1 incident on June 27, 2024. `https://blog.cloudflare.com/cloudflare-1111-incident-on-june-27-2024/`. acessado em 09/02/2025.

[39] T. Holterbach, T. Alfroy, A. D. Phokeer, A. Dainotti, and C. Pelsser. A System to Detect Forged-Origin Hijacks. In *21th USENIX Symposium on Networked Systems Design and Implementation (NSDI 24)*. USENIX Association, 2024.

[40] A. S. Jacobs, R. Beltiukov, W. Willinger, R. A. Ferreira, A. Gupta, and L. Z. Granville. AI/ML for Network Security: The Emperor Has No Clothes. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, CCS '22, page 1537–1551, New York, NY, USA, 2022. Association for Computing Machinery.

[41] P. Kacherginsky. Celer Bridge Incident Analysis. `https://www.coinbase.com/blog/celer-bridge-incident-analysis`. acessado em 29/11/2023.

[42] M. Konte, R. Perdisci, and N. Feamster. ASwatch: An AS Reputation System to Expose Bulletproof Hosting ASes. *SIGCOMM Comput. Commun. Rev.*, 45(4):625–638, aug 2015.

[43] M. Lad, D. Massey, D. Pei, Y. Wu, B. Zhang, and L. Zhang. PHAS: A Prefix Hijack Alert System. In *USENIX Security Symposium*, volume 1, page 3, 2006.

[44] H. Lakkaraju, S. H. Bach, and J. Leskovec. Interpretable Decision Sets: A Joint Framework for Description and Prediction. In *Proc. of the 22nd ACM Discovery and Data Mining*, KDD '16, page 1675–1684, 2016.

[45] Q. Le and T. Mikolov. Distributed Representations of Sentences and Documents. In E. P. Xing and T. Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1188–1196, Bejing, China, 22–24 Jun 2014. PMLR.

[46] M. Lepinski and K. Sriram. BGPsec Protocol Specification. RFC 8205, Sept. 2017.

[47] W. Li, Z. Lin, M. I. Ashiq, E. Aben, R. Fontugne, A. Phokeer, and T. Chung. RoVista: Measuring and Analyzing the Route Origin Validation (ROV) in RPKI. In *Proceedings of the 2023 ACM on Internet Measurement Conference*, IMC '23, page 73–88, New York, NY, USA, 2023. Association for Computing Machinery.

[48] Y. Liu, J. Su, and R. K. Chang. LDC: Detecting BGP Prefix Hijacking by Load Distribution Change. In *2012 IEEE 26th International Parallel and Distributed Processing Symposium Workshops & PhD Forum*, pages 1197–1203, 2012.

[49] K. Lougheed and Y. Rekhter. Border Gateway Protocol (BGP). RFC 1105, June 1989.

[50] R. Lychev, M. Schapira, and S. Goldberg. Rethinking Security for Internet Routing. *Commun. ACM*, 59(10):48–57, sep 2016.

[51] D. Madory. BGP hijack of Twitter by Russian ISP. `https://www.kentik.com/analysis/bgp-hijack-of-twitter-by-russian-isp/`, 2022.

[52] MANRS. About MANRS. `https://manrs.org/about/`. acessado em 02/05/2024.

[53] P. Marcos, L. Prehn, L. Leal, A. Dainotti, A. Feldmann, and M. Barcellos. AS-Path Prepending: There Is No Rose Without a Thorn. In *Proc. ACM IMC´20*, page 506–520, 2020.

[54] Merit Network, Inc. Internet Routing Registry. https://irr.net/, 2024. acessado em 31/07/2024.

[55] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed Representations of Words and Phrases and their Compositionality. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[56] A. Milolidakis, T. Bühler, K. Wang, M. Chiesa, L. Vanbever, and S. Vissicchio. On the Effectiveness of BGP Hijackers That Evade Public Route Collectors. In *IEEE Access*, volume 11, pages 31092–31124, 2023.

[57] A. Milolidakis, T. Bühler, K. Wang, M. Chiesa, L. Vanbever, and S. Vissicchio. On the effectiveness of bgp hijackers that evade public route collectors. *IEEE Access*, 11:31092–31124, 2023.

[58] R. Morillo, J. Furuness, C. Morris, J. Breslin, A. Herzberg, and B. Wang. ROV++: Improved Deployable Defense against BGP Hijacking. In *NDSS*, 2021.

[59] NIST. NIST RPKI Monitor. `https://rpki-monitor.antd.nist.gov/`, 2024.

[60] U. of Oregon. University of Oregon RouteViews Project. https://www.routeviews.org/. acessado em 24/11/2023.

[61] C. Palmeira. Hackers entram na guerra e atacam governos da Palestina e de Israel; entenda. `https://www.tecmundo.com.br/seguranca/272527-hackers-entram-guerra-atacam-governos-palestina-israel.htm`, 2023.

[62] PeeringDB. `https://catalog.caida.org/dataset/peeringdb`, 2010. acessado em 31/07/2024.

[63] Philip Smith. BGP Routing Table Analysis. `https://thyme.apnic.net/`, 2021.

[64] J. Postel. Internet Control Message Protocol. RFC 792, Sept. 1981.

[65] L. Qin, D. Li, R. Li, and K. Wang. Themis: Accelerating the Detection of Route Origin Hijacking by Distinguishing Legitimate and Illegitimate {MOAS}. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 4509–4524, 2022.

[66] R. RCC. Routing Information Service (RIS). https://www.ripe.net/analyse/internet-measurements/routing-information-service-ris. acessado em 24/11/2023.

[67] Y. Rekhter, S. Hares, and T. Li. A Border Gateway Protocol 4 (BGP-4). RFC 4271, Jan. 2006.

[68] M. T. Ribeiro, S. Singh, and C. Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery.

[69] H. Rimlinger, K. Vermeulen, M. Gouel, O. Fourmaux, and T. Friedman. To Probe or Not to Probe? Reconciling High Speed Probing with Ethical Probing. In *Proceedings of the on CoNEXT Student Workshop 2023*, CoNEXT-SW '23, page 19–20, New York, NY, USA, 2023. Association for Computing Machinery.

[70] RIPE NCC RIS. YouTube Hijacking: A RIPE NCC RIS Case Study. `https://www.ripe.net/publications/news/youtube-hijacking-a-ripe-ncc-ris-case-study/`, 2008. acessado em 31/07/2024.

[71] P. Sermpezis, V. Kotronis, A. Dainotti, and X. Dimitropoulos. A Survey among Network Operators on BGP Prefix Hijacking. *SIGCOMM Comput. Commun. Rev.*, 48(1):64–69, apr 2018.

[72] P. Sermpezis, V. Kotronis, P. Gigis, X. Dimitropoulos, D. Cicalese, A. King, and A. Dainotti. ARTEMIS: Neutralizing BGP Hijacking Within a Minute. In *IEEE/ACM Transactions on Networking*, volume 26, pages 2471–2486, 2018.

[73] T. Shapira and Y. Shavitt. A Deep Learning Approach for IP Hijack Detection Based on ASN Embedding. In *Proceedings of the Workshop on Network Meets AI & ML*, NetAI '20, page 35–41, New York, NY, USA, 2020. Association for Computing Machinery.

[74] T. Shapira and Y. Shavitt. AP2Vec: An Unsupervised Approach for BGP Hijacking Detection. *IEEE Transactions on Network and Service Management*, 19(3):2255–2268, 2022.

[75] X. Shi, Y. Xiang, Z. Wang, X. Yin, and J. Wu. Detecting Prefix Hijackings in the Internet with Argus. In *Proceedings of the 2012 Internet Measurement Conference*, IMC '12, page 15–28, New York, NY, USA, 2012. Association for Computing Machinery.

[76] C. A. Shue, A. J. Kalafut, and M. Gupta. Abnormally Malicious Autonomous Systems and Their Internet Connectivity. *IEEE/ACM Transactions on Networking*, 20(1):220–230, 2012.

[77] A. Siddiqui. BGP Security in 2021. `https://manrs.org/2022/02/bgp-security-in-2021/`. acessado em 09/02/2024.

[78] A. Siddiqui. KlaySwap – Another BGP Hijack Targeting Crypto Wallets. `https://www.manrs.org/2022/02/klayswap-another-bgp-hijack-targeting-crypto-wallets/`. acessado em 29/11/2023.

[79] D. N. Silva. História da Internet. `https://brasilescola.uol.com.br/informatica/internet.htm`. acessado em 13/12/2023.

[80] S. Suzuki. A guerra cibernética paralela entre Rússia e Ucrânia. `https://www.bbc.com/portuguese/internacional-60551648`, 2022.

[81] C. Testart, P. Richter, A. King, A. Dainotti, and D. Clark. Profiling BGP Serial Hijackers: Capturing Persistent Misbehavior in the Global Routing Table. In *Proceedings of the Internet Measurement Conference*, IMC '19, page 420–434, New York, NY, USA, 2019. Association for Computing Machinery.

[82] K. Vermeulen, J. P. Rohrer, R. Beverly, O. Fourmaux, and T. Friedman. Diamond-Miner: Comprehensive Discovery of the Internet's Topology Diamonds . In *17th USENIX Symposium on Networked Systems Design and Implementation (NSDI 20)*, pages 479–493, Santa Clara, CA, Feb. 2020. USENIX Association.

[83] W. Willinger, A. Gupta, A. S. Jacobs, R. Beltiukov, R. A. Ferreira, and L. Granville. A NetAI Manifesto (Part I): Less Explorimentation, More Science. *SIGMETRICS Perform. Eval. Rev.*, 51(2):106–108, oct 2023.

[84] Y. Xiang, Z. Wang, X. Yin, and J. Wu. Argus: An Accurate and Agile System to Detecting IP Prefix Hijacking. In *2011 19th IEEE International Conference on Network Protocols*, pages 43–48, 2011.

[85] M. Zeng, X. Huang, P. Zhang, and D. Li. Understanding the Impact of Outsourcing Mitigation Against BGP Prefix Hijacking. *Computer Networks*, 202:108650, 2022.

Apêndice A

Publicações

Durante a realização deste trabalho, os seguintes manuscritos foram produzidos:

- Abrindo a Caixa-Preta – Aplicando IA Explicável para Aprimorar a Detecção de Sequestros de Prefixo. Publicado no Simpósio Brasileiro em Segurança da Informação e de Sistemas Computacionais (SBSeg'24) [20];
- *Uncovering BGP Action Communities and Community Squatters in the Wild.* Publicado em Proceedings of the ACM on Measurement and Analysis of Computing Systems (POMACS) [26] e será apresentado na conferência ACM SIGMETRICS'25 ;
- Caracterização da Vulnerabilidade a Sequestros de Prefixo de Sistemas Autônomos Militares (Submetido) [18].