Cordeiro em Pele de Lobo: Desvelando a Negação de Serviço Baseada em Envenenamento de Reputação

Anderson Frasão[1], Raphael Kaviak Machnicki[1], Tiago Heinrich[2], Vinicius Fulber-Garcia[1]

[1] Universidade Federal do Paraná (UFPR)
Curitiba – PR – Brasil

[2]Max Planck Institute for Informatics (MPI)
Saarbrücken – Saarland – Germany

{aacfrasao,rkmachnicki,vinicius}@inf.ufpr.br,
theinric@mpi-inf.mpg.de

Abstract. *Reputation systems are used to measure the reliability of users, devices and services in digital environments. Although they help with security and decision-making by identifying malicious interactions, these systems are subject to manipulation that can compromise their integrity. This work proposes and validates a new attack vector that exploits reputation systems to carry out denial of service against legitimate users. The attack consists of a malicious agent that impersonates the victim, executes offensive actions and induces automated systems to penalize it based on its reputation. The strategy exploits identity verification flaws in behavior-based trust mechanisms. The attack was demonstrated through experiments with a real service and security system, highlighting its effectiveness in blocking legitimate clients through a triangulated attack and emphasizing the need to explore new methods for detecting and mitigating the proposed attack.*

Resumo. *Sistemas de reputação são utilizados para medir a confiabilidade de usuários, dispositivos e serviços em ambientes digitais. Apesar de seu auxílio na segurança e na tomada de decisão, identificando interações maliciosas, esses sistemas estão sujeitos a manipulações que podem comprometer sua integridade. Este trabalho propõe e valida um novo vetor de ataque que explora sistemas de reputação para realizar negação de serviço contra usuários legítimos específicos. O ataque consiste em um agente malicioso que se passa pela vítima, executa ações maliciosas e induz sistemas automatizados a penalizá-la com base em sua reputação. A estratégia explora falhas na verificação de identidade em mecanismos de confiança baseados em comportamento. O ataque foi demonstrado por meio de experimentos com um serviço e um sistema de segurança reais, evidenciando sua efetividade em bloquear clientes legítimos a partir de um ataque triangulado e destacando a necessidade de explorar novos métodos de detecção e mitigação do ataque proposto.*

1. Introdução

Os sistemas de reputação auxiliam na segurança digital, servindo como mecanismos para avaliar a confiabilidade de usuários, dispositivos ou serviços [Feitosa and Garcia 2016].

Esses sistemas são utilizados em plataformas online como redes sociais até sistemas de autenticação e proteção contra fraudes. A reputação de um usuário é frequentemente medida por meio de avaliações de histórico de atividades ou outros indicadores que podem influenciar no nível ou qualidade de acesso a serviços e interações em um ecossistema digital [Jøsang and Ismail 2002].

Em plataformas de comércio eletrônico, por exemplo, um vendedor com boa reputação tem maior probabilidade de conquistar a confiança dos compradores, enquanto em redes corporativas, um dispositivo confiável pode obter permissões específicas. Esses sistemas também ajudam a filtrar *spam*, detectar comportamento malicioso e mitigar riscos em serviços críticos, como bancos e plataformas de comunicação [Yan et al. 2015,Fulber-Garcia et al. 2018].

Apesar de sua relevância, os sistemas de reputação também apresentam limitações e estão sujeitos a diferentes ameaças. Um dos principais riscos é a manipulação da reputação, que pode ocorrer por meio de falsificação de avaliações, ataques coordenados para inflar ou reduzir a reputação de um cliente, e/ou ataques de personificação. Esses ataques comprometem a confiabilidade do sistema e podem ser explorados para prejudicar alvos específicos [Xu et al. 2015]. Porém, apesar das vulnerabilidades em sistemas de reputação serem conhecidas, poucos são os vetores de ataques bem estruturados e claramente definidos nesse contexto.

Considerando esses aspectos, este artigo propõe um mergulho profundo em uma estratégia ofensiva que envolve personificação para negação de serviço a partir de manipulação de reputação: ataques aqui denominados de **Negação de Serviço Baseada em Envenenamento de Reputação**.

Nesse cenário, um atacante assume a identidade de um cliente legítimo e realiza ações maliciosas deliberadamente para prejudicar sua reputação. Como resultado, o cliente legítimo pode ser bloqueado ou perder o acesso a serviços essenciais. A vulnerabilidade explorada nesse ataque reside na dificuldade de verificação da identidade real por parte do sistema de reputação [Friedman et al. 2007]. Apesar de existirem mecanismos de avaliação baseados no histórico de comportamento, eles não utilizam métodos robustos para garantir que uma ação ou requisição foi realmente executada pelo cliente identificado a partir do tráfego de rede. Dessa forma, ao realizar um ataque sob a identidade de um cliente legítimo, degradando a sua reputação, o atacante pode desencadear punições automáticas que afetam injustamente este cliente, impedindo eventualmente o acesso ao serviço requisitado e o tornando vítima de um ataque triangulado de negação de serviço.

Para demonstrar a execução do ataque de negação de serviço baseado em envenenamento de reputação, um cenário de testes foi construído utilizando um Sistema de Prevenção à Intrusão (IPS) baseado em reputação (Suricata IPS) para proteger um serviço HTTP (Nginx) — ambos bastante conhecidos e amplamente utilizados. Os resultados demonstraram que o ataque proposto foi capaz de indisponibilizar o serviço HTTP, via bloqueio no IPS, para um usuário legítimo, vítima do ataque, por meio da ação maliciosa de degradação de reputação realizada por um atacante.

Sendo assim, pontualmente, destacam-se as principais contribuições do trabalho:

- Proposta e especificação de um novo vetor de ataques relacionado à exploração de vulnerabilidades em sistemas de reputação para a execução de um ataque de

negação de serviço dirigido e triangulado;
- Demonstração prática do ataque proposto utilizando sistemas de reputação e serviços reais.

O restante do artigo está organizado como segue: a Seção 2 apresenta os principais conceitos necessários para a compreensão da proposta. A Seção 3 revisa os trabalhos relacionados. A Seção 4 detalha, tecnicamente, o ataque proposto. A Seção 5 apresenta a metodologia de execução e teste do ataque. A Seção 6 apresenta e discute os resultados obtidos, e a Seção 7 conclui o trabalho.

2. Fundamentação Teórica

Esta seção apresenta a fundamentação teórica deste trabalho, incluindo ataques de negação de serviço (Seção 2.1), ataques de personificação (Seção 2.2) e sistemas de reputação (Seção 2.3).

2.1. Ataques de Negação de Serviço

Os ataques de Negação de Serviço (DoS - Denial of Service) visam tornar um serviço indisponível para usuários legítimos ao sobrecarregar seus recursos computacionais ou explorando vulnerabilidades específicas [Needham 1993].

Existem diferentes categorias de ataques DoS, incluindo os ataques orientados a protocolos específicos e os ataques por inundação de tráfego (*flooding*). Os ataques orientados a protocolos, também conhecidos como ataques semânticos, exploram falhas ou comportamentos específicos de um protocolo ou serviço, como o envio de requisições malformadas que provocam falhas ou travamentos. Esse tipo de ataque requer um conhecimento aprofundado do alvo [Mirkovic et al. 2004]. Por outro lado, os ataques de inundação visam indisponibilizar um serviço ao esgotar sua capacidade de processamento ou largura de banda, por meio do envio massivo de tráfego. Isso pode tornar o sistema inacessível [Jonker et al. 2017].

Uma evolução dos ataques DoS são os ataques de Negação de Serviço Distribuída (DDoS – *Distributed Denial of Service*), nos quais uma rede de dispositivos comprometidos (*botnet*) é utilizada para amplificar a capacidade do ataque. Um exemplo conhecido é a *Mirai Botnet*, que, em 2016, explorou centenas de milhares de dispositivos da Internet das Coisas (IoT) para realizar ataques em larga escala, impactando grandes empresas como Amazon, Netflix e X (antigo Twitter) [Antonakakis et al. 2017].

Os ataques DDoS modernos se tornaram ainda mais sofisticados, empregando técnicas avançadas para evitar detecção e mitigar contra-medidas defensivas. Diferentes abordagens são utilizadas para coordenar ataques, incluindo a arquitetura *Agent-Handler*, onde um atacante controla múltiplos agentes distribuídos (*bots*), e ataques baseados em redes IRC, que dificultam o rastreamento das origens dos comandos. Além disso, ataques podem ser projetados para limitar trafego legitimo, dificultando a detecção por sistemas tradicionais de monitoramento de rede [Bhuyan et al. 2014].

Outra variação importante dos ataques DDoS são os ataques distribuídos de negação de serviço baseados em reflexão (DRDoS - *Distributed Reflection Denial of Service*), que combinam amplificação de tráfego com ofuscação da origem do ataque. Nesses ataques, o tráfego malicioso é enviado a servidores vulneráveis ou mal configurados (os

refletores), que então respondem às vítimas com mensagens maiores do que as requisições originais. Os protocolos mais usados nesses ataques são, em sua maioria, baseados em UDP, como DNS, NTP, CLDAP, SSDP, entre outros [Heinrich et al. 2021]. Essa técnica permite que um atacante, com recursos limitados, gere um volume massivo de dados contra o alvo, ao mesmo tempo, dificultando sua rastreabilidade por meio do *IP spoofing*.

Além disso, os ataques DRDoS têm evoluído para formas ainda mais sofisticadas, como os ataques multiprotocolo, que utilizam simultaneamente diferentes protocolos de amplificação, e os chamados ataque de *carpet bombing*, em que o tráfego refletido é distribuído entre diversos endereços IP de uma mesma sub-rede, dificultando a detecção e a mitigação [Heinrich et al. 2021, Heinrich et al. 2022]. Esses ataques foram observados em larga escala por *honeypots* multiprotocolo como o MP-H, que registrou mais de 1,4 milhão de ataques DRDoS em dois anos, dos quais cerca de 13,8 mil foram multiprotocolo e mais de 3,7% utilizaram *carpet bombing*. Os ataques multiprotocolo se mostraram mais intensos e duradouros que os monoprotocolares, com média de quase o dobro de requisições por ataque e durações maiores [Heinrich et al. 2021].

Embora ataques de DoS, DDoS e DRDoS não comprometam diretamente dados ou sistemas, seus impactos podem ser severos, interrompendo serviços essenciais, como operações empresariais, bancos e provedores de Internet. Além da interrupção, esses ataques têm sido cada vez mais utilizados como forma de extorsão, com criminosos exigindo pagamentos para cessar as ofensivas, e até como instrumentos em disputas geopolíticas, funcionando como verdadeiras armas cibernéticas [Mirkovic et al. 2004].

Para mitigar os riscos associados a ataques de DDoS, diferentes abordagens têm sido consideradas tanto na academia quanto na indústria. Dentre as principais estratégias, destaca-se a filtragem de tráfego por serviços especializados (*scrubbing*), uso de bloqueio remoto (*remote triggered black holing*), desativação de vetores de amplificação (como serviços DNS ou NTP mal configurados) e a promoção de validação de endereços de origem (SAV - *Source Address Validation*), essencial no combate a ataques com IPs falsificados. Além disso, coalizões entre provedores e iniciativas de padronização têm buscado viabilizar práticas colaborativas de defesa e compartilhamento de dados para melhorar a detecção e mitigação de ataques em tempo real [Hiesgen et al. 2024].

2.2. Ataques de Personificação

De modo geral, ataques de personificação ocorrem quando um invasor se faz passar por um usuário legítimo visando obter acesso não autorizado a recursos, realizar ações fraudulentas ou comprometer a integridade dos sistemas. A complexidade dessas ameaças pode variar desde o roubo simples de credenciais até a exploração avançada de vulnerabilidades em protocolos de segurança [Günther 2014].

Por exemplo, nesses ataques, invasores podem utilizar dispositivos como *IMSI Catchers* para se passarem por torres de celular legítimas, interceptando comunicações e rastreando usuários [Park et al. 2019]. Na computação em nuvem, ataques de *session hijacking* permitem que invasores roubem tokens de autenticação para se passarem por usuários legítimos, acessando dados e serviços sensíveis [Thangavel et al. 2017]. Em redes de Wi-Fi, o ataque conhecido como *Evil Twin* consiste na criação de um ponto de acesso falso com o mesmo nome de uma rede confiável, induzindo vítimas a se conectar e expondo suas credenciais [Shrivastava et al. 2020].

O processo de execução de um ataque de personificação se inicia com a escolha de um alvo. Em seguida, o atacante realiza uma pesquisa detalhada sobre a vítima, coletando dados pessoais, comportamentais, profissionais e tecnológicos para construir um perfil convincente. Com base nessas informações, o invasor define um pretexto: uma narrativa estratégica que será utilizada para enganar terceiros. Na etapa seguinte, ocorre a personificação propriamente dita, em que o atacante assume a identidade da vítima por meio de mensagens, documentos falsificados, redes sociais, endereços de sistema ou identificadores clonados. Por fim, o ataque é executado quando o criminoso utiliza essa identidade para realizar transações, obter informações confidenciais ou enganar pessoas e sistemas.

Ataques de personificação podem ocorrer em diversos contextos tecnológicos, cada um com vulnerabilidades e desafios específicos para sua detecção e mitigação. A seguir, destacam-se alguns dos contextos mais relevantes nos quais esse tipo de ataque tem se mostrado particularmente crítico.

Computação em Nuvem No contexto da computação em nuvem, ataques de personificação são especialmente desafiadores, devido à necessidade de correlacionar atividades de um mesmo usuário em diferentes máquinas virtuais e redes distribuídas. Segundo [Kholidy 2021], a detecção desses ataques requer sistemas de monitoramento capazes de identificar anomalias comportamentais. Técnicas como a análise de chamadas de sistema e de fluxo de rede têm sido empregadas para detectar padrões de uso suspeitos e correlacionar eventos potencialmente maliciosos.

Redes Móveis Os ataques de personificação também afetam redes móveis, comprometendo a autenticação mútua entre usuários e a infraestrutura da operadora. Um exemplo é o IMP4GT (IMPersonation in 4G neTworks), um ataque que explora vulnerabilidades no protocolo TLE para permitir que um adversário se passe por um usuário legítimo ou por uma estação base da operadora [Rupprecht et al. 2020]. Diferentemente de ataques anteriores, o IMP4GT não apenas redireciona o tráfego, mas também possibilita a injeção e modificação de pacotes de dados, tornando a identificação e mitigação do ataque um desafio considerável.

Sistemas Bluetooth Dispositivos Bluetooth são suscetíveis a ataques de personificação devido a falhas nos procedimentos de autenticação. O ataque BIAS (Bluetooth Impersonation AttackS), descrito por [Antonioli et al. 2020], explora vulnerabilidades na especificação do protocolo para permitir que um invasor estabeleça conexões seguras sem possuir a chave de autenticação originalmente compartilhada. Esse ataque compromete diretamente a integridade das comunicações Bluetooth e pode afetar diversos dispositivos, como smartphones, laptops e sistemas IoT.

Redes Neurais Uma forma emergente de ataque de personificação envolve o uso de redes neurais generativas para imitar vozes humanas. Avanços recentes com redes adversárias generativas (GANs) permitem a criação de vozes sintetizadas que replicam com alta fidelidade o tom, a cadência e o estilo de fala de indivíduos específicos [Gao et al. 2018]. Esse tipo de ataque representa uma ameaça crescente a sistemas de autenticação baseados em biometria de voz, exigindo contramedidas robustas, como a análise de padrões fonéticos e espectrogramas para validar a autenticidade da fala.

Os ataques de personificação configuram um desafio multifacetado que compromete diversos setores da segurança da informação. Desde a exploração de vulnerabilida-

des em protocolos de comunicação, até o uso de inteligência artificial para falsificação de identidade, esses ataques exigem o desenvolvimento de técnicas avançadas de detecção e mitigação. Estratégias como correlação de auditorias, análise de tráfego e uso de aprendizado de máquina são fundamentais para aprimorar a segurança dos sistemas contra essa classe de ameaças.

2.3. Sistemas de Reputação

Os sistemas de reputação surgiram como um mecanismo essencial para estabelecer confiança em ambientes distribuídos, onde as interações frequentemente ocorrem entre partes desconhecidas ou anonimas. Esses sistemas agregam e analisam *feedbacks* de múltiplos usuários para gerar confiabilidade e mitigar riscos associados à tomada de decisões em ambientes digitais [Hendrikx et al. 2015].

Tais sistemas podem ser classificados em duas grandes categorias: (i) centralizados, nos quais uma autoridade única coleta e processa as avaliações dos participantes, disponibilizando-as publicamente; e (ii) distribuídos, que eliminam um ponto central de controle e propagam avaliações entre os participantes de uma rede. O modelo centralizado é amplamente empregado em plataformas de comércio eletrônico, como eBay e Amazon, onde as classificações dos usuários influenciam diretamente a confiabilidade percebida de vendedores e compradores [Jøsang et al. 2007]. Já os sistemas distribuídos são frequentemente adotados em redes *peer-to-peer (P2P)* e na cibersegurança para avaliar a confiabilidade de nós em um ambiente descentralizado [Jøsang et al. 2007].

Em cenários de tomada de decisão em grupo, a reputação - entendida como a percepção coletiva da confiabilidade de uma pessoa com base em avaliações anteriores - desempenha um papel fundamental na filtragem de especialistas que fornecem avaliações confiáveis. Modelos baseados em reputação ajudam a construir confiança entre especialistas e a incentivar comportamentos cooperativos ao longo do tempo. Por exemplo, no modelo proposto por [You et al. 2024], a reputação é calculada considerando tanto o *feedback* direto (avaliação de interações anteriores) quanto recomendações indiretas (opiniões de terceiros). Esse tipo de abordagem visa minimizar a influência de agentes maliciosos e garantir que a confiança refletida nos especialistas seja representativa de sua credibilidade real.

Além do cenário eletrônico e de tomada de decisões, sistemas de reputação são utilizados na defesa contra ataques baseados em DNS e redes de sensores sem fio. Em redes de sensores, modelos como BTRES (*Beta-based Trust and Reputation Evaluation System*) são utilizados para monitorar e avaliar a confiabilidade dos nós, permitindo a detecção e mitigação de ataques internos [Fang et al. 2016]. Por exemplo, em uma rede de sensores ambientais em uma área remota, o BTRES pode detectar um nó comprometido, evitando leituras falsas de temperatura ao atribuir a ele uma reputação negativa, com base no desvio em relação aos dados dos demais sensores. Da mesma forma, sistemas baseados em listas de bloqueio (*blocklist*) são empregados para identificar e bloquear domínios maliciosos em serviços como DNS e listas de *e-mails* [Sinha et al. 2008]. Um exemplo típico é o uso do Spamhaus, que mantém listas de IPs e domínios associados a envio de spam; servidores de e-mail consultam essas listas para recusar mensagens de remetentes maliciosos. No cenário de tecnologias emergentes, sistemas de reputação compõem mitigadores de ataques DDoS, como o apresentado no serviço DeMONS [Fulber-Garcia et al. 2018], baseado no paradigma de virtualização de funções de rede.

Entre tanto, pesquisas recentes indicam que esses sistemas podem ser vulneráveis a ataques estratégicos, como manipulação de popularidade, permitindo que domínios maliciosos evitem detecção ao simular comportamento benigno, esses sistemas também sofrem com ataques *Sybil* (criação de identidades falsas) e a limitação da propagação de confiança em redes complexas [Galloway et al. 2024].

No contexto de cibersegurança, ataques direcionados demonstraram que é possível evadir sistemas de reputação, comprometendo sua eficácia em bloquear ameaças. [Galloway et al. 2024] evidenciaram que atacantes podem modificar suas estratégias para manipular métricas de reputação e, com um investimento financeiro relativamente baixo, conseguir burlar sistemas de reputação baseados em DNS com taxa de evasão de até 100% para as soluções testadas.

Além disso, o uso de *feedbacks* apresenta algumas limitações importantes. Primeiramente, há o risco de avaliações tendenciosas, especialmente em contextos de tomada de decisão em grupo, nos quais especialistas podem agir por autopromoção ou com a intenção de prejudicar concorrentes. Outra limitação é a dificuldade em garantir a confiabilidade dessas avaliações, sejam elas provenientes de seres humanos ou de sistemas automatizados de análise [You et al. 2024]. Assim, embora sistemas de reputação componham muitas arquiteturas e soluções de segurança, ainda há uma série de desafios a serem superados para que as reputações calculadas sejam plenamente confiáveis.

3. Trabalhos Relacionados

Ataques de negação de serviço orientados a ações de personificação para envenenamento e deterioração da credibilidade representam um desafio emergente para sistemas de reputação e controle de acesso. Essa ameaça é especialmente relevante em plataformas que utilizam reputação como critério de acesso, como sistemas de *e-commerce*, redes sociais e ambientes distribuídos baseados em *feedback* dos usuários. No entanto, no melhor de nosso conhecimento, o estado da arte ainda não aborda explicitamente essa categoria de ataque de negação de serviço, uma vez que ela apresenta etapas, características e objetivos bastante específicos em sua execução. Por outro lado, vulnerabilidades em sistemas de reputação e ataques pontuais de personificação explorando essas vulnerabilidades têm sido marginalmente discutidos na literatura.

Em [Etesami et al. 2016], os autores investigam a influência do conformismo e da manipulação em sistemas de reputação, utilizando um modelo baseado em teoria dos jogos para demonstrar como agentes podem distorcer avaliações para influenciar a percepção pública. Embora o estudo forneça ideias valiosas sobre a dinâmica de opiniões, não aborda explicitamente ataques de personificação ou estratégias que visem à negação de serviço por meio do envenenamento e deterioração da reputação.

[Xiong et al. 2007] propõem um *framework* para mitigar a escassez e a manipulação de *feedback* em sistemas de reputação, utilizando técnicas de inferência baseadas em similaridade. O trabalho investiga como essas técnicas podem aumentar a resiliência dos sistemas contra manipulações. No entanto, não considera a possibilidade de ataques que explorem o envenenamento e a deterioração da reputação como forma de negação de serviço, como também não discute técnicas de personificação como um potencial vetor de ataque.

[Friedman et al. 2007] exploram vulnerabilidades em sistemas de reputação,

incluindo ataques de falsificação de identidade e geração de *feedbacks* artificiais (*sybil attacks*). O estudo discute contramedidas para ataques que envolvem *feedbacks* fraudulentos e analisa a resistência de sistemas de reputação contra manipulações estratégicas, propondo modelos matemáticos para avaliar o impacto de diferentes ataques. Apesar de abordar a manipulação de reputação e a criação de identidades falsas, o trabalho não discute ataques que introduzem negação de serviço por meio de envenenamento e deterioração da reputação, especialmente aqueles que envolvem a personificação de usuários legítimos.

[Xu et al. 2015] investigam a manipulação de reputação em plataformas de *e-commerce*, detalhando a operação de mercados clandestinos que fornecem serviços de escalonamento de reputação (*Reputation-Escalation-as-a-Service*). O estudo analisa a criação de *feedbacks* falsos e a comercialização de identidades para inflacionar artificialmente a credibilidade de vendedores, demonstrando a sofisticação das estratégias de manipulação. No entanto, os autores não abordam ataques que visam o envenenamento e deterioração da reputação de usuários legítimos como forma de negação de serviço.

Os artigos revisados fornecem uma base teórica relevante para a compreensão da manipulação de reputação, mas apresentam limitações que evidenciam lacunas específicas. Em particular, nenhum deles aborda diretamente o problema de ataques de personificação voltados à negação de serviço por meio do envenenamento e deterioração da reputação. Essa ausência na literatura reforça a necessidade de investigações sobre os traços estruturais desse tipo de ataque, de modo a viabilizar, posteriormente, a proposição de contramedidas efetivas. Este trabalho busca preencher essa lacuna ao investigar um modelo de ataque específico, contribuindo para a compreensão e caracterização da negação de serviço baseada em envenenamento de reputação.

4. Ataque de Negação de Serviço Baseada em Envenenamento de Reputação

Este trabalho propõe e específica os ataques de negação de serviço baseada em envenenamento de reputação, combinando técnicas de personificação e manipulação de sistemas de reputação para negar recursos a uma vítima benigna e legítima. Embora a literatura trate separadamente ataques de personificação e falhas em sistemas de reputação, a combinação desses vetores em um ataque direcionado à manipulação da reputação de uma entidade específica ainda é pouco explorada, sendo, até onde sabemos, inexistente a discussão acadêmica sobre o uso dessa abordagem como forma de ataque de negação de serviço.

No contexto geral de envenenamento de sistemas de reputação, considera-se um modelo de ameaça onde o atacante objetiva manipular a reputação de uma vítima legítima. Particularmente, para a negação de serviço, essa reputação deve ser degradada. Assim, o atacante se passa pela vítima (personificação) e realiza ações maliciosas detectáveis, levando o sistema de reputação a associar tais comportamentos à vítima. Como consequência, a vítima sofre punições, como o bloqueio de acesso aos serviços protegidos por esses sistemas, caracterizando uma forma específica de negação de serviço.

A Figura 1 ilustra o ataque. Inicialmente, existe a fase de personificação, na qual o atacante assume a identidade da vítima. Isso pode se feito de várias maneiras, dependendo do serviço-alvo escolhido. Por exemplo, o atacante pode roubar credenciais de autenticação da vítima (como senhas, *tokens* ou chaves de *API*) por meio de *phishing*,

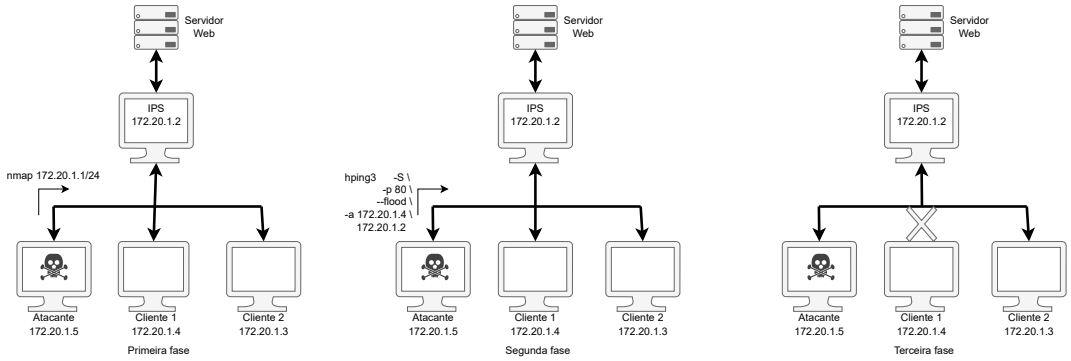

Figura 1. Representação do ataque de Negação de Serviço por Envenenamento de Reputação.

engenharia social ou exploração de vulnerabilidades. Alternativamente, o atacante pode falsificar informações de identificação, como endereços IP, endereços de *e-mail* ou certificados digitais, para se passar pela vítima. Ainda, em sistemas com mecanismos de autenticação fracos, o atacante pode explorar brechas para assumir a identidade da vítima sem precisar de suas credenciais completas [Esparza 2019].

O atacante utiliza, então, a identidade roubada para realizar ações maliciosas contra o serviço-alvo. Essas ações podem variar dependendo do tipo de serviço e do objetivo do atacante. No caso específico de degradação de reputação, o atacante pode, por exemplo, enviar um grande volume de requisições ao serviço, sobrecarregando-o e levando o provedor a identificar a vítima como a origem de um possível ataque de negação de serviço (DoS). Outra possibilidade é a realização de atividades que violem os termos de uso do serviço-alvo, como o envio de spam [Google 2025], a postagem de conteúdo inadequado ou a tentativa de exploração de vulnerabilidades. Em alguns casos, o atacante pode ainda alterar ou corromper dados associados à vítima, afetando diretamente sua reputação no sistema.

Com as atividades maliciosas em andamento, o provedor do serviço-alvo detecta padrões compatíveis com ataques e adota medidas para proteger seu sistema. Como o atacante personifica a vítima, o provedor associa as ações maliciosas à identidade da vítima legítima. Isso pode resultar na suspensão ou bloqueio da conta da vítima, na redução de sua reputação no sistema ou até mesmo em sua inclusão em uma lista de bloqueio. Assim, as medidas de segurança, embora projetadas para proteger o serviço-alvo, acabam negando o acesso de uma vítima benigna e legítima como resultado de um ataque triangulado.

Ressalta-se que, para a execução de ataques de negação de serviço baseado em envenenamento de reputação, a etapa de reconhecimento da vítima e do serviço-alvo é fundamental. Além da identificação de informações relacionadas à vítima, necessária para uma personificação eficaz, é também imprescindível mapear as características, políticas e sistemas de segurança do serviço-alvo – neste caso, não para evitá-los, mas para explorá-los ao máximo, gerando alertas abundantes e provocando o bloqueio da vítima pelo próprio serviço-alvo, como consequência da ação do atacante.

O impacto na vítima pode ser significativo e vai além da simples indisponibilidade do serviço. A degradação da reputação pode ter consequências a longo prazo, especial-

mente em sistemas onde esta é crucial para a participação, como *marketplaces online*, redes sociais ou sistemas de colaboração. Além disso, a vítima pode sofrer perdas financeiras diretas (por exemplo, em serviços de *e-commerce*) ou indiretas (por exemplo, interrupção de operações criticas). Reverter o bloqueio ou restaurar a reputação pode ser um processo demorado e complexo, exigindo que a vítima prove sua inocência ao provedor do serviço.

5. Metodologia de Avaliação

Para demonstrar a execução do ataque proposto, foi estabelecido um ambiente de teste executando sistemas de segurança reais para a análise da detecção de comportamentos maliciosos e a resposta automatizada a eventos de ataque. O cenário experimental utilizou quatro contêineres: (i) um servidor web baseado na imagem oficial do Nginx [1]; (ii) o sistema Suricata, em modo IPS, com suporte a reputação de IPs, utilizando a imagem mantida por Jason Ish [2], também executada em contêiner Docker; (iii) um cliente benigno e legítimo, vítima do ataque, construído a partir da imagem oficial do Debian; e (iv) um atacante, construído a partir da imagem oficial do Debian, responsável por realizar o ataque ao servidor. Todas as requisições feitas ao servidor, sejam do cliente legítimo ou do atacante, são avaliadas pelo IPS, que as encaminha ao servidor Nginx ou as descarta conforme a reputação do requerente.

Os experimentos foram realizados em um ambiente Linux 6.12.13 utilizando a distribuição Linux Debian 12, o ambiente de virtualização escolhido para os testes consiste no Docker 20.10.24.

Para a comunicação, foram configuradas duas sub-redes em modo *bridge* (definindo os IPs de forma estática). A primeira sub-rede foi utilizada para a comunicação entre o servidor web e o Suricata; já a segunda, para a comunicação entre o Suricata e outras máquinas, neste caso, o cliente legítimo e o atacante. Para a realização do ataque, basta que o atacante identifique o endereço IP do cliente legítimo que será vítima do ataque, possibilitando a personificação.

Após o atacante identificar o IP da vítima (cliente legítimo), utilizando a ferramenta `hping3`, foi realizado um *flood* de pacotes SYN para o servidor. Com isso, o *Suricata* identifica o ataque e reduz progressivamente a reputação do cliente, até o ponto em que ele é bloqueado – é importante notar que, em momentos intermediários, antes do bloqueio total, o IPS descarta uma parcela das requisições recebidas, baseado na reputação atual do cliente. Durante os testes, o cliente vítima mantinha um tráfego contínuo de requisições legítimas ao servidor, permitindo monitorar a disponibilidade do serviço e o tempo de resposta. A interrupção na recepção dessas respostas foi o principal indicador de que o sistema de reputação havia acionado o bloqueio do cliente vítima.

As instruções completas para reprodução do ambiente, assim como os códigos e arquivos de configuração utilizados, estão disponíveis em um repositório público no GitHub[3], permitindo a replicação dos testes e a verificação independente dos resultados.

Para ilustrar um cenário de ataque, considere que um atacante deseja impedir que

[1] `https://hub.docker.com/_/nginx`

[2] `https://hub.docker.com/r/jasonish/suricata/`

[3] `https://github.com/Carmofrasao/SBSeg-2025-Frasao`

uma empresa utilize um serviço de nuvem. O atacante rouba as credenciais de um funcionário da empresa e as utiliza para enviar um grande volume de requisições maliciosas ao serviço de nuvem. Ao detectar o tráfego anormal, o provedor de nuvem bloqueia o endereço IP associado à empresa, impedindo que todos os funcionários acessem o serviço. A empresa, agora vítima, sofre degradações e interrupções no serviço e, consequentemente, em suas operações, necessitando investir tempo e recursos para resolver o problema com o provedor.

Outro exemplo prático consiste no cenário onde o atacante deseja que uma vítima não receba *e-mails* de um serviço. Assim, o atacante se passa pelo usuário legítimo usando técnicas de *spoofing* para falsificar o endereço do remetente [Babu et al. 2010]. O atacante envia um grande volume de *spam* para o serviço alvo, um serviço *online* que depende de um sistema de reputação para identificar atividades maliciosas. O serviço, ao detectar o comportamento intrusivo associado ao endereço da vítima, reduz a reputação e, eventualmente, bloqueia o usuário para proteger o sistema. Como resultado, o usuário legítimo deixa de receber *e-mails*, sofrendo uma negação de serviço. Para recuperar o acesso, a vítima precisa provar que foi personificada e solicitar o desbloqueio de sua conta, um processo que pode ser demorado e complexo.

6. Avaliação Experimental

Os experimentos realizados permitiram observar na prática o comportamento do ataque proposto, caracterizado por uma estratégia de personificação maliciosa visando provocar uma negação de serviço direcionada e triangulada. O modelo explora mecanismos de reputação automatizada, como os implementados pelo Suricata, para induzir o bloqueio de um cliente legítimo, sem que este realize diretamente qualquer ação maliciosa.

A dinâmica do ataque parte do conhecimento de que o sistema de monitoramento vincula o comportamento de rede ao endereço IP de origem. Assim, o atacante personifica a vítima pelo seu IP, direcionando tráfego malicioso ao servidor. Como o sistema de reputação detecta esse padrão anômalo, ele penaliza o IP associado, ou seja, o da vítima, levando ao seu bloqueio. O resultado é uma indisponibilidade forçada, causada não pela vítima, mas por quem se passou por ela.

Para avaliar essa dinâmica, dois indicadores principais foram observados: a **quantidade de pacotes enviados** e o **tempo de resposta das requisições**. Ambos os valores foram registrados ao longo do tempo, com destaque para os momentos antes, durante e após os ataques de personificação/*flood*, degradando a reputação da vítima. A expectativa era observar um aumento gradual no volume de pacotes e, paralelamente, uma degradação no tempo de resposta até o ponto de interrupção total da conexão do cliente legítimo.

A Figura 2 demonstra a execução do ataque de negação de serviço baseado em envenenamento, conforme descrito anteriormente. A linha contínua com círculos pretos representa o fluxo de requisições legítimas feitas pelo cliente ao servidor. A linha tracejada com quadrados em cinza-escuro mostra o fluxo de requisições maliciosas, feitas pelo atacante personificando o cliente legítimo, e direcionadas ao servidor. A linha tracejada e pontilhada com triângulos em cinza-claro indica o encaminhamento das requisições recebidas, sejam elas legítimas ou não, do Suricata (IPS) ao servidor, para efetiva resposta.

O fluxo de requisições maliciosas, ataque de *flooding*, ocorre no período entre 30s e 306s, alcançando um pico de 111 pacotes de sincronização enviados em janelas de 10

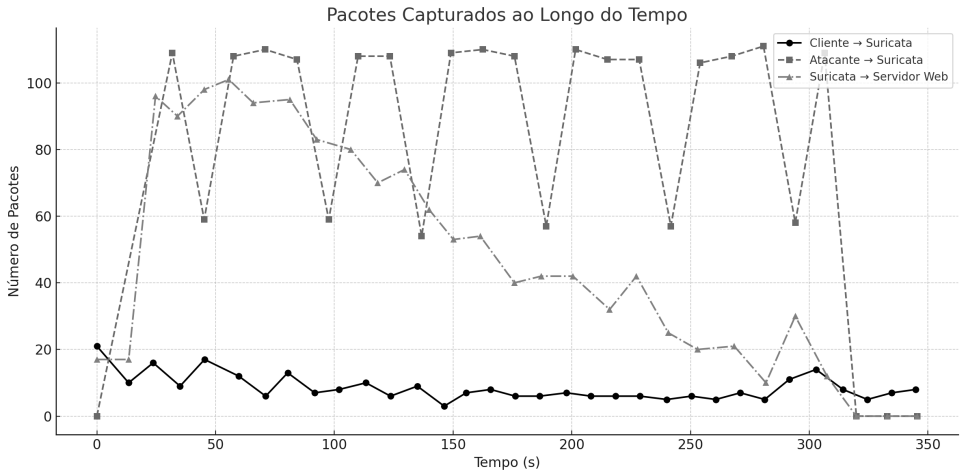

Figura 2. Evolução do volume de pacotes gerados na rede.

Figura 3. Variação do tempo de resposta percebido pelo cliente legítimo ao fazer requisições para o servidor

segundos. A partir de aproximadamente 319s, o atacante encerra o fluxo de requisições maliciosas, restando somente o cliente enviando requisições com, em média, 10 pacotes. Entretanto, este último não consegue mais se comunicar com o servidor, já que está bloqueado no IPS, confirmando que o ataque de negação de serviço baseado em envenenamento de reputação foi bem-sucedido. Ressalta-se que a repetição controlada desses ataques visa simular um comportamento que ultrapassa os limiares definidos nas regras do Suricata, acionando o mecanismo de penalidade por reputação.

Já o gráfico de tempo de resposta às requisições legítimas, presente na Figura 3, apresenta um padrão de aumento contínuo, refletindo a degradação gradual no desempenho da comunicação entre o cliente legítimo e o servidor. Inicialmente, os tempos de resposta variam de forma relativamente estável, com picos pontuais. No entanto, a partir de aproximadamente 70 segundos, observa-se um aumento abrupto e progressivo no tempo de resposta, ocasionado pela aplicação das políticas de descarte parcial do IPS, culminando em um tempo extremo de mais de 16 segundos na última requisição. A partir deste ponto, nenhuma requisição legítima é respondida, uma vez que há o descarte total

dos fluxos de requisição do cliente legítimo pelo IPS. Esse crescimento, acompanhado da queda abrupta no tráfego de resposta (Figura 2), confirma o momento em que a vítima foi de fato bloqueada.

Esses resultados reforçam a premissa de que sistemas baseados em reputação, embora eficazes em contextos tradicionais de detecção, estão sujeitos à manipulação em cenários de personificação. A falta de validação, da identidade real de quem origina os pacotes, possibilita a exploração desse tipo de falha lógica no processo de decisão. Vale destacar que o ataque não requer um volume alto de tráfego para ter sucesso. O foco está na persistência e na capacidade de simular com fidelidade o padrão que leva o sistema a reconhecer uma origem como maliciosa. Dessa forma, trata-se de uma ameaça mais sutil e, potencialmente, mais perigosa do que ataques convencionais de negação de serviço, especialmente quando existe um alvo específico a ter o serviço negado.

A exploração de sistemas baseados em reputação, como demonstrados nos testes com o Suricata, revela uma vulnerabilidade que pode afetar uma gama de soluções de segurança automatizadas, incluindo firewalls, sistemas de prevenção de intrusões (IPS) e mecanismos de mitigação em CDNs. Qualquer arquitetura que baseie decisões de bloqueio exclusivamente no endereço IP de origem, sem validação contextual ou autenticação adicional, pode ser manipulado por técnicas similares. O impacto para indivíduos é considerável: clientes legítimos podem ser excluídos de serviços essenciais sem qualquer ação maliciosa de sua parte. Em ambientes corporativos, isso pode significar a interrupção de fluxos de comunicação, perda de produtividade e até danos à credibilidade. Já em contextos sensíveis e críticos como serviços digitais de governo ou planos de saúde, o bloqueio indevido de clientes pode ter consequências ainda mais graves.

Por fim, os experimentos confirmam a viabilidade do modelo proposto de **ataques de negação de serviço baseada em envenenamento de reputação**, evidenciando uma fragilidade nos mecanismos de reputação em sistemas de segurança que pode ser explorada para causar negação de serviço triangulada em alvos específicos.

7. Conclusão

Os sistemas de reputação são fundamentais para estabelecer confiança em ambientes digitais, desde plataformas de comércio eletrônico até redes corporativas. No entanto, sua dependência de indicadores comportamentais e a falta de mecanismos robustos de verificação de identidade os tornam vulneráveis a manipulações maliciosas. Este artigo investigou um novo ataque que explora essas fragilidades: os Ataques de Negação de Serviço Baseada em Envenenamento de Reputação, em que um atacante se passa por uma cliente legítimo (vítima do ataque) para degradar sua reputação e, assim, provocar sua exclusão ou bloqueio por sistemas de segurança automatizados.

Por meio de uma abordagem teórica e experimental, foi demonstrado como a personificação combinada com ações maliciosas pode ser instrumentalizada para prejudicar clientes legítimos. Os experimentos, realizados em um ambiente controlado com ferramentas como Suricata e técnicas como *spoofing* de IP, confirmaram que ataques persistentes — mesmo com volume baixo/moderado de tráfego — podem enganar sistemas de reputação, resultando no bloqueio injusto da vítima. Os dados coletados revelam padrões claros: o aumento gradual do tempo de resposta e o eventual bloqueio total do tráfego de requisições legítimas, evidenciando o sucesso da estratégia de negação de serviço por

envenenamento de reputação.

As implicações deste trabalho devem ser estudadas mais a fundo por exporem uma fragilidade crítica em sistemas de segurança baseados em reputação amplamente conhecidos. O primeiro passo, a partir do que foi apresentado neste artigo, deve ser a identificação de métodos para detectar ataques de envenenamento de reputação, independentemente de sua finalidade; entre as técnicas potencialmente relevantes, pode-se incluir a análise comportamental histórica do cliente e a verificação de traços de rota do tráfego de rede. A partir da detecção, estratégias de mitigação podem ser desenvolvidas, buscando separar e bloquear efetivamente o tráfego malicioso personificado, enquanto se mantém o atendimento às requisições legítimas, garantindo a continuidade do serviço para um cliente vítima de um ataque de envenenamento de reputação. Para isso, os traços de detecção podem ser utilizados em classificadores leves, eficientes e eficazes que auxiliem na tomada de decisão dos sistemas de prevenção à intrusão e de outros filtros sofisticados de tráfego de rede.

Em trabalhos futuros, propõe-se integrar mecanismos adicionais de validação de identidade (como assinaturas digitais [Rakhra et al. 2024] ou analise contextual de dispositivos [Sae-Bae and Memon 2014]) e a revisão de limiares de reputação para evitar falsos positivos. Como direção futura, destacamos a necessidade de testar o ataque em cenários mais complexos (como autenticação multifatorial ou ambientes de nuvem) e desenvolver ferramentas adaptativas que correlacionam reputação a múltiplas dimensões de confiança.

Agradecimentos

Este trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES PROEX), Centro de Computação Científica e Software Livre (C3SL) e Fundação de Âmparo à Pesquisa e Inovação de Santa Catarina (FAPESC). Os autores também agradecem o Programa de Pós-Graduação em Informática da Universidade Federal do Paraná.

Referências

Antonakakis, M. et al. (2017). Understanding the mirai botnet. In *Security Symposium*, pages 1093–1110.

Antonioli, D. et al. (2020). Bias: Bluetooth impersonation attacks. In *Symposium on Security and Privacy*, pages 549–562. IEEE.

Babu, P. R. et al. (2010). A comprehensive analysis of spoofing. *International Journal of Advanced Computer Science and Applications*, 1(6).

Bhuyan, M. H. et al. (2014). Detecting distributed denial of service attacks: methods, tools and future directions. *The Computer Journal*, 57(4):537–556.

Esparza, J. M. (2019). Understanding the credential theft lifecycle. *Computer Fraud & Security*, 2019(2):6–9.

Etesami, S. R. et al. (2016). Conformity versus manipulation in reputation systems. In *Conference on Decision and Control*, pages 4451–4456. IEEE.

Fang, W. et al. (2016). Btres: Beta-based trust and reputation evaluation system for wireless sensor networks. *Journal of Network and Computer Applications*, 59:88–94.

Feitosa, D. d. L. and Garcia, L. S. (2016). Sistemas de reputação: um estudo sobre confiança e reputação no comércio eletrônico brasileiro. *Revista de Administração Contemporânea*, 20(1):84–105.

Friedman, E. et al. (2007). Manipulation-resistant reputation systems. *Algorithmic Game Theory*, 677.

Fulber-Garcia, V. et al. (2018). Demons: A ddos mitigation nfv solution. In *International Conference on Advanced Information Networking and Applications*, pages 769–776. IEEE.

Galloway, T. et al. (2024). Practical attacks against dns reputation systems. In *Symposium on Security and Privacy*, pages 4516–4534. IEEE.

Gao, Y. et al. (2018). Voice impersonation using generative adversarial networks. In *International Conference on Acoustics, Speech and Signal Processing*, pages 2506–2510. IEEE.

Google (2025). Políticas de spam para a Pesquisa Google na Web. Acessado em 19 de abril de 2025.

Günther, C. (2014). A survey of spoofing and counter-measures. *NAVIGATION: Journal of the Institute of Navigation*, 61(3):159–177.

Heinrich, T. et al. (2021). New kids on the drdos block: Characterizing multiprotocol and carpet bombing attacks. In *International Conference on Passive and Active Network Measurement*, pages 269–283. Springer.

Heinrich, T. et al. (2022). Um estudo de correlaçao de ataques drdos com fatores externos visando dados de honeypots. In *Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais (SBSeg)*, pages 358–371. SBC.

Hendrikx, F. et al. (2015). Reputation systems: A survey and taxonomy. *Journal of Parallel and Distributed Computing*, 75:184–197.

Hiesgen, R. et al. (2024). The age of ddoscovery: an empirical comparison of industry and academic ddos assessments. In *Internet Measurement Conference*, pages 259–279. ACM.

Jonker, M. et al. (2017). Millions of targets under attack: a macroscopic characterization of the dos ecosystem. In *Internet Measurement Conference*, pages 100–113.

Jøsang, A. et al. (2007). A survey of trust and reputation systems for online service provision. *Decision support systems*, 43(2):618–644.

Jøsang, A. and Ismail, R. (2002). The beta reputation system. In *Bled Electronic Commerce Conference*, volume 160, pages 324–337.

Kholidy, H. A. (2021). Detecting impersonation attacks in cloud computing environments using a centric user profiling approach. *Future Generation Computer Systems*, 117:299–320.

Mirkovic, J. et al. (2004). *Internet denial of service: attack and defense mechanisms (Radia Perlman Computer Networking and Security)*. Prentice Hall PTR.

Needham, R. M. (1993). Denial of service. In *Conference on Computer and Communications Security*, pages 151–153. ACM.

Park, S. et al. (2019). Anatomy of commercial imsi catchers and detectors. In *Workshop on Privacy in the Electronic Society*, pages 74–86. ACM.

Rakhra, M. et al. (2024). Digital signature verification in cloud computing. In *International Conference on Reliability, Infocom Technologies and Optimization*, pages 1–6. IEEE.

Rupprecht, D. et al. (2020). Imp4gt: Impersonation attacks in 4g networks. In *Network and Distributed System Security Symposium*. The Internet Society.

Sae-Bae, N. and Memon, N. (2014). Online signature verification on mobile devices. *Transactions on Information Forensics and Security*, 9(6):933–947.

Shrivastava, P. et al. (2020). Evilscout: Detection and mitigation of evil twin attack in sdn enabled wifi. *Transactions on Network and Service Management*, 17(1):89–102.

Sinha, S. et al. (2008). Shades of grey: On the effectiveness of reputation-based "blacklists". In *International Conference on Malicious and Unwanted Software*, pages 57–64.

Thangavel, M. et al. (2017). Session hijacking over cloud environment: A literature survey. *Advancing Cloud Database Systems and Capacity Planning With Dynamic Applications*, pages 363–391.

Xiong, L. et al. (2007). Countering feedback sparsity and manipulation in reputation systems. In *International Conference on Collaborative Computing: Networking, Applications and Worksharing*, pages 203–212. IEEE.

Xu, H. et al. (2015). E-commerce reputation manipulation: The emergence of reputation-escalation-as-a-service. In *International Conference on World Wide Web*, pages 1296–1306.

Yan, S.-R. et al. (2015). A graph-based comprehensive reputation model: Exploiting the social context of opinions to enhance trust in social commerce. *Information Sciences*, 318:51–72.

You, X. et al. (2024). A reputation-based trust evaluation model in group decision-making framework. *Information Fusion*, 103:102082.