The "Sheep in Wolf's Clothing" Attack: Denial-of-Service via Reputation Poisoning

Anderson Frasão, Raphael K. Machnicki
Federal University of Paraná
Paraná, Brazil
{aacfrasao, rkmachnicki}@inf.ufpr.br

Tiago Heinrich
Max Planck Institute for Informatics
Saarbrücken, Germany
theinric@mpi-inf.mpg.de

Vinicius Fulber-Garcia
Federal University of Paraná
Paraná, Brazil
vinicius@inf.ufpr.br

***Abstract*—Reputation systems are commonly used to assess the reliability of users, devices, and services. Although they support security and decision-making by identifying malicious activities, such systems are themselves susceptible to manipulation. This paper presents and validates a novel attack, called Sheep in Wolf's Clothing (SWC). In this attack, an adversary impersonates a victim and executes detectable malicious actions against services protected by reputation-based network security solutions, thereby poisoning the victim's reputation, prompting these solutions to unfairly penalize the victim, and ultimately denying the victim access to the service. The effectiveness of the SWC attack is demonstrated through a case study using a widely known Intrusion Prevention System (IPS).**

***Index Terms*—Reputation, Poisoning, Impersonation, DoS.**

I. Introduction

Reputation systems support digital security by assessing the reliability of users, devices, or services [1], and are commonly integrated into social networks, authentication mechanisms, and network security solutions. An entity's reputation is typically derived from its activity history, which, for example, determines the level, priority, or quality of access it is granted to requested resources [2]. Thus, reputation systems facilitate operations such as spam filtering, malicious behavior detection, and risk mitigation in critical services.

Despite their relevance, reputation systems also present limitations and are vulnerable to attacks. One significant threat is reputation poisoning, which may occur through fake reviews (*i.e.*, coordinated efforts to inflate or deflate an entity's reputation artificially) or impersonation attacks. Such attacks undermine the reliability of reputation systems and can be exploited to harm specific victims [3]. Although these vulnerabilities are well explored in some contexts, such as e-commerce, they remain underexplored in network security. As a result, naive reputation systems, susceptible to threats that fly under the radar, are being deployed alongside state-of-the-practice network security solutions.

Considering the described scenario, this paper presents a novel impersonation-and-poisoning attack that exploits vulnerabilities in reputation systems used in network security, resulting in the denial of a legitimate service to a legitimate victim (*i.e.*, a reputation-based denial-of-service). We refer to this as the **Sheep in Wolf's Clothing (SWC)** attack.

In an SWC attack, an adversary impersonates a victim and deliberately executes detectable malicious actions to gradually degrade the victim's reputation within a service protected by a reputation-based network security solution. The attack stems from the inherent complexity of reliably confirming identities in network traffic [4]: although reputation mechanisms based on activity history are commonly employed, they often lack robust guarantees that the observed activities were indeed performed by the entity identified in the traffic. As a result, executing the SWC attack allows the adversary to trigger automatic penalties that unjustly affect the victim, ultimately leading to a triangulated denial-of-service.

To demonstrate the effectiveness of the SWC attack, we conducted a case study in which we deployed the Suricata IPS, a popular Intrusion Prevention System, with its native reputation module to protect an NGINX HTTP service. In this scenario, an adversary impersonates a legitimate user (the victim, who continuously issues requests to the HTTP service) and executes the SWC attack. The results indicated that the attack was successful, gradually degrading the victim's reputation and finally rendering the HTTP service unavailable. These outcomes motivated a discussion of the attack's limitations, challenges, and potential implications for the future of reputation systems.

The rest of this paper is organized as follows: Section II presents relevant background concepts. Section III technically details the SWC attack. Section IV presents the case study, results, and discussion. Section V indicates and briefly reviews related works on reputation manipulation. Finally, Section VI concludes the paper.

II. Background

This section briefly presents the fundamentals of denial-of-service attacks (Section II-A), impersonation attacks (Section II-B), and reputation systems (Section II-C).

A. Denial-of-Service Attacks

Denial-of-Service (DoS) attacks aim to render a resource, such as an online service, unavailable to legitimate users [5]. Based on this general definition, a variety of techniques for carrying out DoS attacks have been identified over the years [6]. For instance, a DoS can be executed through protocol-specific or traffic-flooding attacks. Protocol-specific attacks exploit the specific behaviors of a given protocol or service, sending requests that cause crashes or degrade resources. In contrast, traffic-flooding attacks seek to exhaust the resource's

processing capacity or available bandwidth by generating a massive volume of requests.

A particular variation of DoS is the Distributed Denial of Service (DDoS) attack, in which a network of compromised devices is used to amplify the attack's capacity [7]. DDoS attacks have become increasingly sophisticated, employing advanced strategies to evade detection and defensive countermeasures [8]. Examples include the agent–handler architecture, in which an adversary controls multiple distributed malicious agents, and attacks coordinated via Internet Relay Chat (IRC) networks, which hinder the tracing of command origins.

A third variation is the Distributed Reflection Denial-of-Service (DRDoS) attack, which combines traffic amplification with obfuscation of the attack source [9], [10]. In such attacks, malicious traffic is directed to vulnerable or misconfigured servers (the reflectors), which subsequently send responses to the victim that are significantly larger in volume than the original requests.

Anyway, regardless of the specific variation or technique, a denial-of-service attack is fundamentally characterized by its objective: denying legitimate users access to a legitimate resource. Naturally, detection, mitigation, and prevention mechanisms aim to identify and contain such attacks by blocking adversarial entities. As a consequence, adversaries commonly spoof their identities to evade these security solutions [8].

B. Impersonation Attacks

Impersonation attacks involve an adversary posing as a legitimate user in order to gain unauthorized access to resources, execute fraudulent actions, or compromise system integrity [11]. Typically, a successful impersonation attack involves multiple execution stages [12]: (i) the adversary selects a victim; (ii) the adversary builds a victim profile by gathering the information necessary for impersonation; (iii) the adversary defines a pretext, *i.e.*, the actions and narrative to be presented to third parties during the attack; (iv) the adversary impersonates the victim and initiates the attack; and (v) the adversary obtains the targeted resources or performs the desired actions while assuming the victim's identity, leveraging the victim's privileges and redirecting the consequences to them.

Impersonation attacks can occur in a variety of technological contexts. For example, in telecommunications, adversaries can use devices such as IMSI Catchers to impersonate legitimate cell towers, intercept communications, and track users [13]. In cloud computing, session-hijacking attacks allow adversaries to steal authentication tokens and impersonate legitimate users, thereby accessing sensitive data and services [14]. In Wi-Fi networks, the Evil Twin attack involves creating a fake access point with the same name as a trusted network, inducing victims to connect and exposing their credentials [15].

Overall, impersonation typically constitutes a first-stage attack that enables an adversary to execute more complex multi-stage attacks or to exfiltrate targeted resources. Detecting and mitigating such attacks is nontrivial and often requires sophisticated solutions employing, for example, behavior analytics and machine learning [16].

C. Reputation Systems

Reputation systems aim to establish trust in distributed environments, where interactions frequently occur among unknown or anonymous entities. These systems aggregate and analyze information from multiple sources and/or apply predefined policies to assess an entity's reliability, thereby enabling proactive actions to mitigate imminent or potential threats based on the outcomes [17].

Reputation systems generally follow two main models [18]: (i) centralized, in which a single authority processes information and policies to determine the reputations of entities; and (ii) distributed, which eliminates centralized management and operates cooperatively, with multiple authorities processing and sharing reputation data. Based on these models, reputation systems have been applied in several contexts, such as monitoring and assessing the reliability of sensor nodes [19], identifying malicious domains [20], and detecting DDoS-related malicious traffic flows [21].

However, recent works indicate that reputation systems may be vulnerable to different attacks, such as reputation poisoning, typically aiming to increase the trust of malicious entities by simulating benign behavior or faking information about them, Sybil attacks, with the creation and manipulation of false identities, and limited propagation of trust in complex networks [22], [23]. Another limitation concerns the difficulty of ensuring the reliability of reputation assessments, whether derived from human input or automated analysis systems [24]. Consequently, although reputation systems are widely incorporated into cybersecurity solutions, several open challenges remain.

III. Sheep in Wolf's Clothing Attack

This paper presents and characterizes a novel denial-of-service attack focused on reputation systems that combines impersonation and reputation poisoning to deny a benign, legitimate victim access to a targeted resource. Although prior work has explored variants of impersonation attacks and reputation poisoning, their integrated execution to deliberately manipulate reputation remains unexplored.

As previously discussed, reputation poisoning refers to attacks that artificially manipulate an entity's reputation. Typical reputation-poisoning attacks aim to inflate an entity's reputation to increase its trust level, performance, or access privileges. In contrast, when reputation poisoning is leveraged to mount a denial-of-service attack, the victim's reputation must be significantly degraded within the system. To achieve this, an adversary impersonates the victim and executes detectable malicious actions, leading the reputation system to associate that behavior with the victim. Consequently, the victim is subjected to penalties, such as reduced service quality and, ultimately, denied access to services protected by the reputation system. We refer to the described high-level attack process as the Sheep in Wolf's Clothing attack; details on its adversary model and workflow are presented in Sections III-A and III-B, respectively.

A. Adversary Model

For an SWC attack, we consider an adversary with the following characteristics:

Objective. The adversary seeks to compromise the availability of a specific resource for a benign victim by inducing a reputation-based security solution to misclassify the victim as a malicious entity.

Capabilities. The adversary must be capable of acting as the victim. For example, it must be able to forge network traffic that impersonates the victim, including packets carrying well-known malicious signatures, and transmit such traffic to simulate malicious behavior.

Knowledge. The adversary must be aware that a reputation-based security solution is deployed and actively protecting the target resource. Furthermore, although not mandatory, the adversary should be aware of the malicious patterns (*e.g.*, signatures and behaviors) identified by the security solution. Finally, the adversary must previously know or actively discover the characteristics of the victim (*e.g.*, addresses, identifiers, and names) that will be impersonated.

Limitations. The adversary cannot modify the policies or analysis procedures employed by the reputation system, nor does the adversary have direct means to assess the success of the attack, as the victim experiences the collateral effects. However, the adversary may exploit indirect means to infer the attack's effectiveness.In the context of these experiments, the attacker was on the internal network. However, if the attacker is on an external network, there is a possibility that they could be blocked due to measures such as RFC 2827, which blocks spoofed traffic originating from external networks [25].

Attack Scope. This adversary model encompasses impersonation, reputation poisoning, and denial-of-service attacks.

B. Attack Workflow

The general workflow of the SWC attack is illustrated in Figure 1. Initially, the attack enters the impersonation phase, during which the adversary assumes the victim's identity. This phase can be carried out in several ways, depending on the targets: the victim and the resource to be denied. For instance, the adversary may steal the victim's authentication credentials (*e.g.*, passwords, tokens, or API keys) through phishing, social engineering, or vulnerability exploitation [26]. Alternatively, the adversary may forge identifying information, such as IP addresses, email addresses, or digital certificates, to impersonate the victim [27].

In the second stage, reputation poisoning, the adversary uses the stolen identity to perform malicious actions against the target resource. These actions may vary depending on the operation of the reputation-based security solution and the adversary's objectives. For instance, the adversary may generate a large volume of requests to a service, overloading it and leading the security solution to identify the victim as the source of a potential denial-of-service attack. Alternatively, the adversary may engage in activities that violate the target resource's terms of use, such as sending spam emails,

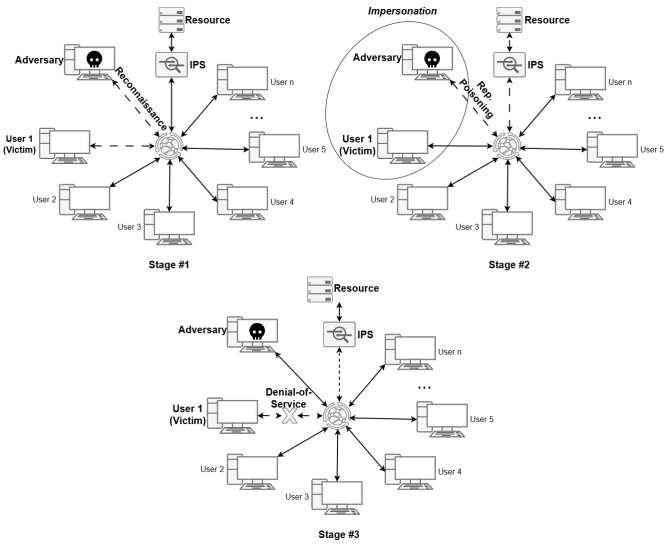

Fig. 1. SWC Attack Workflow

posting inappropriate content, or attempting to exploit known vulnerabilities. In some cases, the adversary may also modify, or corrupt data associated with the victim, thereby directly degrading the victim's reputation within the system.

Given the ongoing malicious activity, the reputation-based security solution protecting the target resource is expected to detect patterns consistent with attacks and to take appropriate preventive or mitigating actions. Because the adversary impersonates the victim, the service provider associates the malicious behavior with the victim's identity, thereby penalizing the victim's reputation. Such penalties may lead to a degradation of the victim's privileges and, ultimately, to their permanent blocking from accessing the protected resource. In this scenario, the security solution effectively detects and mitigates malicious activity against the target resource; however, rather than constraining the actual adversary, it denies access to the impersonated victim, constituting the third stage of a SWC attack: denial-of-service.

It should be highlighted that accurately identifying both the victim and the target resource is fundamental to conducting SWC attacks. Beyond gathering information about the victim, which is necessary for effective impersonation, it is also essential to understand and map the characteristics, policies, and mechanisms of the reputation-based security solutions protecting the target resource. In this context, the goal is not to evade these mechanisms but rather to deliberately exploit them, generating numerous alerts and causing the victim to be penalized and blocked as quickly as possible.

Although denying service is the primary objective of an SWC attack, its consequences may extend beyond the unavailability of a resource. Reputation poisoning can have long-term effects, particularly in scenarios where reputation is central to participation, such as online marketplaces, social networks, and collaborative or distributed systems. In addition, the victim may incur direct losses (for example, in e-commerce environments) or indirect financial losses (for example, due

to the interruption of critical operations caused by resource unavailability) [28]. Furthermore, restoring access to a resource or restoring reputation can be a long and complex process, often requiring the victims to demonstrate their integrity to the resource provider and managers.

IV. Case Study: Demonstrating SWC

To demonstrate the SWC attack in practice, a case study was designed using a real-world, widely used reputation-based security solution to detect malicious behavior and automate the prevention and mitigation of potential attacks: the Suricata [1] Intrusion Prevention System (IPS).

In summary, the testing scenario consists of four entities, deployed as containers: (i) a web server NGINX[2], as the target resource; (ii) the Suricata IPS, with IP reputation support enabled[3], as the reputation-based security solution protecting the resource; (iii) a benign client, the victim of the attack, that legitimately requests the resource; and (iv) an adversary, responsible for carrying out the SWC attack. All requests made to the server, whether from the legitimate client, the victim, or an adversary impersonating the victim, are evaluated by the IPS, which forwards them to the NGINX server or discards them based on the request's origin (source IP) reputation.

All tests were conducted in a Linux 6.12.13 environment using the Debian 12 distribution. The virtualization environment was based on Docker 20.10.24. Two bridge-mode subnets were configured to enable communication among the entities. The first subnet was dedicated to communication between the web server and the IPS (the restricted network). On the other hand, the second subnet allowed the client and the adversary to access the resource (the web server) after undergoing IPS analysis (the public network). In particular, the client attempts to issue a legitimate GET request to the server every 10 seconds.

To carry out the SWC attack in this case study, we assume the adversary knows the victim's IP address, enabling impersonation through forged network traffic. The adversary, impersonating the victim, sends packets with signatures and in patterns that emulate three well-known attacks, possibly identified by the IPS: a command-and-control attack based on CVE2007-2663[4]; a SYN flood attack, sending 100 synchronization packets in sequence; and an HTTP flood attack, sending 100 POST packets in sequence.

The IPS rule set includes signatures for command-and-control attacks and patterns for SYN flood attacks, enabling the identification of such attacks. However, no rule is defined to detect HTTP flood attacks. Each new source is initially considered fully trusted (with a reputation score of 384 points) and experiences a reputation degradation for every attack alert generated by the IPS (losing 20 points per alert). The reputation score determines the packet-pass and packet-dropping rates applied to traffic from a known source, as defined by the policy in Equations 1 and 2.

$$\text{PassRate}_{\text{Source}} = \frac{\text{Reputation}_{\text{Source}}}{384} \tag{1}$$

$$\text{DropRate}_{\text{Source}} = 1 - \text{PassRate}_{\text{Source}} \tag{2}$$

The case study comprises the following phases: (i) during the first 30 seconds, the victim legitimately accesses the web server, and no attack is in progress; (ii) after this initial period, the SWC attack begins, and the adversary, impersonating the victim, conducts repeated attack rounds following a fixed pattern. Specifically, within a 10-second interval, command-and-control packets are sent; in the subsequent 10-second interval, SYN flood packets are generated; finally, in a third 10-second interval, an HTTP flood is launched. The adversary repeats these rounds as necessary until the IPS denies the victim access to the web server. During the tests, only the victim and the adversary interact with the IPS and the web server. It is important to note that Suricata evaluates all the incoming packets by default, before they being sent to the web server.

The victim continued to issue legitimate requests to the web server throughout the attack. To evaluate the impact of reputation degradation, we measured the mean Round-Trip Time (RTT) observed by the victim when receiving responses to legitimate requests. The RTT is inherently influenced by the traffic policing enforced by the IPS and becomes unmeasurable (*i.e.*, effectively infinite) once the service is denied to the victim. Scripts, tools, and detailed technical instructions for reproducing the case study are available in a public GitHub repository[5], enabling replication of the experiments and independent verification of the results.

Figure 2 presents the lifecycle and the effects related to the execution of the SWC attack, as described in this section. Figure 2-a indicates the number of packets associated with the client/victim IP address, regardless of whether they are legitimate or spoofed, that pass through the IPS and reach the web server. Figure 2-b demonstrates the number of packets sent by the adversary impersonating the victim that arrive at the IPS. This traffic shows naive and well-known command-and-control, SYN flood, and HTTP flood attack patterns. Figure 2-c presents the number of legitimate packets sent by the client that are effectively processed by the IPS, regardless of being forwarded to the web server or not. Finally, Figure 2-d shows the RTT of a legitimate client request that is successfully responded to by the web server (the "X" in the figure indicates that, from that point onward, the client no longer receives any response from the server).

As shown in Figure 2, the SWC attack unfolds gradually: while the poisoning process is ongoing, the victim experiences communication degradation with the web server, ranging from mild to severe, and eventually has the requested service denied. During the first 30 seconds, only the legitimate client (the victim) communicates with the service. The Suricata IPS classifies the victim's IP as benign (reputation 384) and, as no attack alerts are raised, forwards the requests to the server.

[1] https://suricata.io
[2] https://tinyurl.com/DocNginx
[3] https://tinyurl.com/DocSuricata
[4] https://tinyurl.com/2007-2663
[5] https://tinyurl.com/SWC-Test

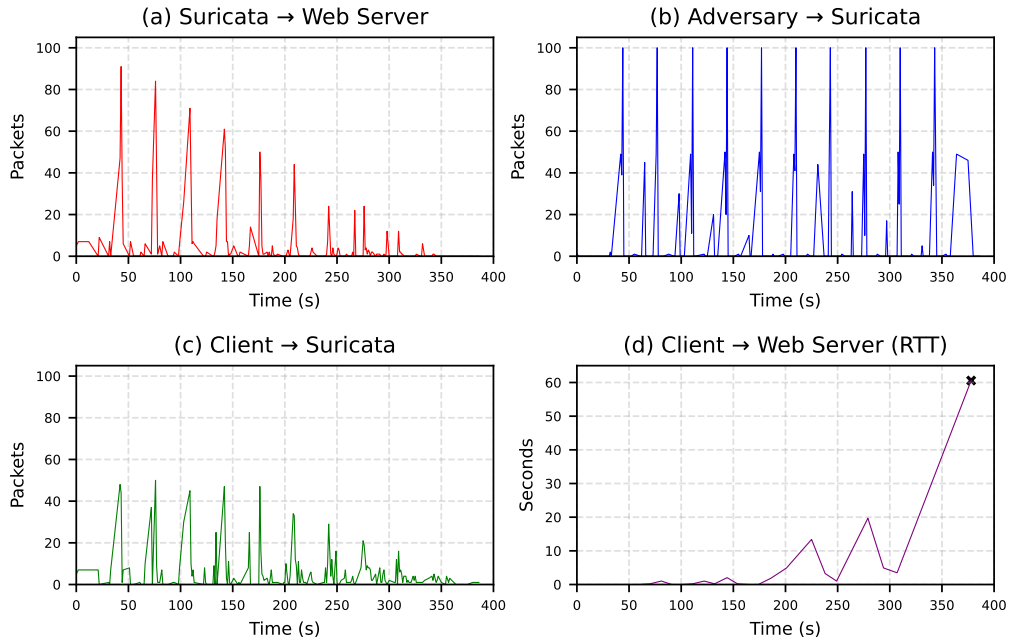

Fig. 2. SWC Attack Execution: Lifecycle and Effects

After 30 seconds, the adversary impersonates the victim and begins executing naïve attacks against the web server. As the IPS has detection rules for two of the executed attacks (Command-and-Control and SYN flood), the victim's reputation begins to decline. Due to IPS traffic policing, the reduced reputation results in partial drops in network traffic sourced from the victim's IP address. This behavior can be observed in Figure 2-a, which shows a progressive reduction in traffic sourced from the victim's IP address being forwarded by the IPS to the web server, and in Figure 2-c, which shows a progressive increase in the RTT experienced by the victim when receiving responses to legitimate requests from the web server. Finally, after 390 seconds, the victim is denied access to the service.

An interesting phenomenon can be observed in Figure 2-c: while the adversary is carrying out the SWC attack, the number of packets originating from the legitimate client that reach the IPS initially peaks and then gradually decreases. This behavior occurs because the adversary executes a SYN flood attack while impersonating the victim, prompting the web server to respond with SYN-ACK packets. These SYN-ACK packets are delivered to the legitimate client, which did not initiate the connections, and therefore responds with RST packets. As the client's reputation decreases and the IPS increasingly blocks the SYN flood traffic, the number of RST packets generated by the legitimate client also decreases, eventually dropping to zero when the service is fully denied, and the SWC attack ceases. Notably, even unknowingly, the client/victim may ultimately contribute to the SWC attack by generating responses that would not have occurred otherwise, thereby increasing the server's load. Note that, in other contexts, an adversary could strategically exploit this characteristic to leverage the victim as a reflector of malicious actions in an SWC attack.

Based on the presented results, we confirm that the SWC attack was successfully executed, achieving its objective of denying service to a specific client/victim by poisoning their reputation. A key aspect of SWC is its ability to deceive security solutions into believing that threats have been effectively mitigated by blocking malicious sources. It occurs because many security solutions, particularly network-based ones, currently lack robust mechanisms to confirm the true identity of traffic sources, thereby expanding the attack surface for SWC attacks.

V. Related Works

Reputation poisoning and impersonation are emerging challenges. These threats have been studied in the context of e-commerce platforms, social networks, and collaborative services that rely on feedback, with discussions focusing on specific vulnerabilities in reputation systems and on impersonation attacks that exploit them for various purposes.

Etesami et al. [29], investigate the influence of artificial manipulation of reputation systems, using a game-theoretic model to demonstrate how agents can distort ratings to shape public perception. Although the study provides valuable insights into opinion dynamics, it does not explicitly address impersonation attacks or denial-of-service strategies through reputation poisoning.

Xiong et al. [30] propose a framework to mitigate scarcity and manipulation of feedback in reputation systems by leveraging similarity-based inference techniques. However, it does not consider the possibility of attacks exploiting reputation poisoning to deny access to a resource, nor the use of impersonation techniques as a potential threat to reputation systems.

Friedman et al. [4] investigates vulnerabilities in reputation systems, including identity spoofing attacks and the generation

of artificial feedback (Sybil attacks). The study discusses countermeasures against fraudulent feedback attacks, analyzes the resistance of reputation systems to strategic poisoning, and proposes mathematical models to assess the impact of various attacks. Although it addresses reputation poisoning and the creation of false identities, the work does not discuss them in the context of denial-of-service attacks, especially those involving the impersonation of legitimate users.

Xu et al. [3] investigate reputation poisoning on e-commerce platforms, detailing the operation of underground markets that provide reputation escalation services (Reputation-Escalation-as-a-Service). The study analyzes the creation of fake feedback and the commercialization of identities to inflate sellers' credibility artificially. However, they do not address attacks that aim to poison reputation through impersonation of legitimate users.

The related works primarily focus on establishing the fundamentals of reputation poisoning and presenting representative attack scenarios. However, none of them directly consider the potential for manipulating a reputation system through impersonation to enable a targeted denial-of-service attack. This work fills this gap by introducing the SWC attack, detailing its adversarial model and workflow, and laying the groundwork for future discussions on its prevention, detection, and mitigation.

VI. Conclusion

This paper presents the Sheep in Wolf's Clothing (SWC) attack, which leverages reputation poisoning and impersonation to deny a specific victim access to a resource. The attack is described and technically validated through a case study using a widely adopted intrusion prevention system (Suricata) with its native reputation module protecting an NGINX web server. The results indicate that vulnerabilities in reputation systems can be successfully exploited, leading to service denial for a legitimate client/victim.

Our findings demonstrate that the SWC attack poses a tangible threat, highlighting the need for effective prevention, detection, and mitigation strategies. Potential directions for future work include distinguishing legitimate sources from spoofed ones through digital signatures and blueprints [31], as well as employing contextual device analysis [32].

ACKNOWLEDGEMENTS

This research was partially financed by CAPES PROEX (88881.189840/2025-01), CNPq (444820/2024-8), UFPR (23075.058047/2025-51) and C3SL - Center for Scientific Computing and Free Software.

References

[1] O. Hasan, L. Brunie, and E. Bertino, "Privacy-preserving reputation systems based on blockchain and other cryptographic building blocks: A survey," *ACM Computing Surveys*, vol. 55, 2022.

[2] A. Jøsang and R. Ismail, "The beta reputation system," in *Bled Electronic Commerce Conf.*, vol. 160, 2002.

[3] H. Xu *et al.*, "E-commerce reputation manipulation: The emergence of reputation-escalation-as-a-service," in *Int. Conf. on World Wide Web*, 2015.

[4] E. Friedman *et al.*, "Manipulation-resistant reputation systems," *Algorithmic Game Theory*, vol. 677, 2007.

[5] S. Ramanauskaite and A. Cenys, "Taxonomy of dos attacks and their countermeasures," *Central European Journal of Computer Science*, vol. 1, 2011.

[6] U. Tariq *et al.*, "A comprehensive categorization of ddos attack and ddos defense techniques," in *Int. Conf. on Advanced Data Mining and Applications*, 2006.

[7] B. L. Dalmazo *et al.*, "A systematic review on distributed denial of service attack defense mechanisms in programmable networks," *Int. Journal of Network Management*, vol. 31, 2021.

[8] M. H. Bhuyan *et al.*, "Detecting distributed denial of service attacks: methods, tools and future directions," *The Computer Journal*, vol. 57, 2014.

[9] T. Heinrich *et al.*, "New kids on the drdos block: Characterizing multiprotocol and carpet bombing attacks," in *Int. Conf. on Passive and Active Network Measurement*, 2021.

[10] T. Heinrich, N. C. Will, R. R. Obelheiro, and C. A. Maziero, "Anywhere on earth: A look at regional characteristics of drdos attacks." in *ICISSP*, 2024, pp. 21–29.

[11] C. Günther, "A survey of spoofing and counter-measures," *NAVIGATION: Journal of the Institute of Navigation*, vol. 61, 2014.

[12] M. Campobasso and L. Allodi, "Impersonation-as-a-service: Characterizing the emerging criminal infrastructure for user impersonation at scale," in *Proceedings of the 2020 ACM SIGSAC Conf. on computer and communications security*, 2020.

[13] C. Yu *et al.*, "Lte phone ignorenumber catcher: A practical attack against mobile privacy," *Security and Communication Networks*, vol. 2019, 2019.

[14] S. Wedman *et al.*, "An analytical study of web application session management mechanisms and http session hijacking attacks," *Information Security Journal: A Global Perspective*, vol. 22, 2013.

[15] P. Shrivastava *et al.*, "Evilscout: Detection and mitigation of evil twin attack in sdn enabled wifi," *Trans. on Network and Service Management*, vol. 17, 2020.

[16] A. Sharma *et al.*, "Mitigating social engineering attacks through ai and behavioral analytics in human-centric cybersecurity," in *Int. Conf. on Data Analytics & Management*, 2025.

[17] F. Hendrikx *et al.*, "Reputation systems: A survey and taxonomy," *Journal of Parallel and Distributed Computing*, vol. 75, 2015.

[18] A. Jøsang *et al.*, "A survey of trust and reputation systems for online service provision," *Decision support systems*, vol. 43, 2007.

[19] W. Fang *et al.*, "Btres: Beta-based trust and reputation evaluation system for wireless sensor networks," *Journal of Network and Computer Applications*, vol. 59, 2016.

[20] S. Sinha *et al.*, "Shades of grey: On the effectiveness of reputation-based "blacklists"," in *Int. Conf. on Malicious and Unwanted Software*, 2008.

[21] V. Fulber-Garcia *et al.*, "Demons: A ddos mitigation nfv solution," in *Int. Conf. on Advanced Information Networking and Applications*, 2018.

[22] V. Agate *et al.*, "A simulation software for the evaluation of vulnerabilities in reputation management systems," *ACM Trans. on Computer Systems*, vol. 37, 2021.

[23] T. Galloway *et al.*, "Practical attacks against dns reputation systems," in *Symposium on Security and Privacy*, 2024.

[24] X. You *et al.*, "A reputation-based trust evaluation model in group decision-making framework," *Information Fusion*, vol. 103, 2024.

[25] D. Senie and P. Ferguson, "Network Ingress Filtering: Defeating Denial of Service Attacks which employ IP Source Address Spoofing," RFC 2827, May 2000. [Online]. Available: https://www.rfc-editor.org/info/rfc2827

[26] K. Thomas *et al.*, "Data breaches, phishing, or malware? understanding the risks of stolen credentials," in *ACM SIGSAC Conf. on Computer and Communications Security*, 2017.

[27] M. Conti, N. Dragoni, and V. Lesyk, "A survey of man in the middle attacks," *IEEE communications surveys & tutorials*, vol. 18, 2016.

[28] J.-P. Louisot, "Managing risk to reputation," in *Enterprise Risk Management in Today's World: A Current and Futuristic View of the Complexity, Resilience, Responsibility and Tools in ERM, Part B*. Emerald, 2024.

[29] S. R. Etesami *et al.*, "Conformity versus manipulation in reputation systems," in *Conf. on Decision and Control*, 2016.

[30] L. Xiong *et al.*, "Countering feedback sparsity and manipulation in reputation systems," in *Int. Conf. on Collaborative Computing: Networking, Applications and Worksharing*, 2007.

[31] M. Rakhra *et al.*, "Digital signature verification in cloud computing," in *Int. Conf. on Reliability, Infocom Technologies and Optimization*, 2024.

[32] N. Sae-Bae and N. Memon, "Online signature verification on mobile devices," *Trans. on Information Forensics and Security*, vol. 9, 2014.