Dynamic Analysis Framework for Classifying Malicious

Web Pages

**Abstract**

In the era of growing digital interconnectivity, the need for robust cybersecurity mechanisms has never been more pressing. This research delves into a unique method for malicious URL (malicious website) detection by leveraging browser automation to capture the contextual behavior of URLs, specifically focusing on their interactions with Browser APIs. Traditional methods of malicious URL detection have relied on lexical analysis, content examination, or blacklists. However, these methods, although effective to some extent, are not comprehensive in the face of evolving cyber threats.

Recent global statistics highlight an increase in cyber threats, underscoring the importance of innovative solutions to address vulnerabilities. Our research approach is informed by these recent trends and offers a more holistic view of URL behavior. By observing how URLs interact when visited by a browser, we capture intricate behavioral patterns that may indicate malicious intent.

This technique of context-based URL analysis contrasts with more conventional approaches. The dataset, sourced from widely recognized repositories, comprises both benign and malicious URLs, labeled based on their interactions. Subsequent machine learning models trained on this dataset exhibits promising results.

As the threat landscape continues to shift and adapt, so too must our methodologies, with this research presenting a step in that direction.

**Introduction**

One of the most critical issues in today's digital landscape concerns the presence of malevolent URLs. These URLs serve as gateways to phishing sites and malware downloads, among a wider spectrum of cyberattacks that could jeopardize personal, corporate, and even national security.

To address the problem of harmful URLs, a holistic approach is required that surpasses conventional tactics like blacklists, semantic analysis, and content-based approaches. In the past, methods such as blacklists, lexical analysis, and content-based techniques were utilized to combat this issue, but now more sophisticated approaches are employed. While these traditional mechanisms have provided a certain level of defense, they also have inherent limitations. For instance, blacklists can only target known threats, rendering them less effective against newly generated malevolent URLs. However, lexical analysis may fail to detect URLs that are cleverly designed to evade detection, and content-based methods can be ineffective against dynamic content and require substantial resources.

Accordingly, the central purpose of this analysis is to suggest and clarify a recent method for identifying malicious web sites. Unlike conventional methodologies, this approach incorporates browser automation to capture the contextual behavior of URLs, with a focus on their interactions with Browser APIs, to detect malicious URLs. This research is based on the hypothesis that by monitoring and analyzing URL behavior in a controlled browsing environment, a more accurate understanding of their benign or malicious nature can be attained.

**Research Problems**

The primary research problem is the identification and classification of URLs as benign or malicious. With the exponential growth in the number of websites and the dynamic nature of the internet, manually flagging malicious URLs becomes a herculean task. Furthermore, the modus operandi of cyber adversaries continually evolves, rendering many traditional detection mechanisms obsolete or less effective over time.

**Shortcomings of Existing Approaches for Malicious URL Detection:**

Blacklists: Blacklists, repositories of known malicious URLs, are instrumental in addressing known threats. However, their inherent reactive nature means they consistently lag, unable to detect new malicious URLs as they emerge. The inability to preemptively identify threats before they are widely recognized limits the effectiveness of blacklists [1].

Lexical Analysis: Lexical analysis, focusing on the structure of URLs to identify malicious patterns, can overlook URLs designed to evade such detection. Furthermore, while lexical features like the bag-of-words model are common, they might not always capture the semantic essence of the URL, making detection less effective [2].

Content-Based Analysis: This approach may not be effective in distinguishing phishing websites from benign websites, as phishing websites often have similar content to the authentic websites they target [3].

The limitations of these conventional methods emphasize the necessity for innovative and comprehensive solutions, especially given the constantly evolving nature of cyber threats [4] [5].

**Proposed Method**

The core of this research is a novel approach that diverges from traditional paradigms. Instead of relying solely on the lexical or content attributes of URLs, this method employs browser automation to observe the behavior of URLs in a controlled environment, with a particular emphasis on their interactions with Browser APIs. The premise is simple yet profound: by understanding how URLs behave when accessed by a browser, one can glean insights into their nature - benign or malicious.

This context-centric approach offers several advantages:

Holistic Analysis: Unlike lexical analysis, which might miss intricately crafted malicious URLs [6], or content analysis, which is resource-intensive, the proposed method offers a balanced approach, capturing behavioral patterns indicative of malicious intent.

Adaptability: Given its reliance on behavioral patterns, this method is inherently more adaptable to the evolving tactics employed by cyber adversaries.

**Contributions of the Thesis**

1. Implementation Overview: In tackling the issues of identifying malicious URLs, this investigation explains a rigorous technique based on browser automation methods. Instead of solely depending on the static attributes of URLs, the proposed framework observes and records real-time interactions between URLs and various browser APIs. Such a strategy makes it possible to extract dynamic behavioral characteristics, providing deeper insight into the potential malicious nature of a URL.

2. Datasets: The effectiveness of this research is heavily reliant on a curated dataset that combines benign and malicious URLs. Sourced from reputable repositories, the dataset serves as a testament to the diverse landscape of URLs and plays a key role in modeling intricate patterns associated with malicious web entities. This comprehensive dataset, enhanced with behavioral nuances, forms the foundation for all subsequent analytical endeavors.

3. Experiments and Performance: After utilizing the dataset, machine learning models were trained and then subjected to rigorous evaluations. Early findings have been positive, showcasing the viability of the proposed approach.

**Summary and Organization of the Rest of the Thesis**

This thesis primarily revolves around the innovative approach of malicious URL detection through browser automation. Following this introduction, the ensuing chapters delve deeper into the specifics of the methodology, the dataset's detailed analysis, experimental setups, results, and a comprehensive discussion juxtaposing the proposed method with traditional mechanisms. The concluding sections encapsulate the findings, their implications in the broader realm of cybersecurity, and potential avenues for future research.

**REFERENCES**

[1] Sushant Sinha, Michael Bailey, and Farnam Jahanian. 2008. Shades of Grey: On the effectiveness of reputation-based“blacklists". In MALWARE 2008. IEEE.

[2] Malicious URL Detection based on Machine Learning

Cho Do Xuan, Hoa Dinh Nguyen1, Tisenko Victor Nikolaevich.

[3] Detecting Malicious URLs Based on Machine Learning Algorithms and Word Embeddings.

[4] Malicious URL Detection using Machine Learning: A Survey. Doyen Sahoo, Chenghao Liu, Steven Hoi.

[5] Dharmaraj Rajaram Patil and JB Patil. 2015. Survey on Malicious Web Pages Detection Techniques. International

Journal of u-and e-Service, Science and Technology (2015).

[6] DETECTION OF ALGORITHMICALLY GENERATED MALICIOUS DOMAIN

Enoch Agyepong1, William J. Buchanan2 and Kevin Jones3