The Effectiveness of Anti-Phishing Tools on Perturbed Logos

Faith Chernowski, Reagan Sanz, Radhika Garg
Department of Computer Science, University of Tennessee
Knoxville, TN, USA
Email: {fchernowski, rsanz, rgarg}@utk.edu

***Abstract*—Phishing attacks are evolving to bypass advanced detection systems, with adversaries increasingly using imperceptible perturbations in logos to evade anti-phishing tools. This study investigates the robustness of these tools against such perturbed logos by generating noisy versions of common brand logos and testing their detectability across various anti-phishing systems. We explore the effectiveness of tools like Phish.AI and LogoGuard in recognizing perturbed logos and propose improvements to enhance their detection accuracy. Our research contributes actionable insights toward improving anti-phishing defenses, particularly in visual similarity detection.**

I. Introduction and Motivation

Phishing attacks continue to be a major source of cyber threats, with logos often used to deceive users into believing they are interacting with legitimate brands. This study aims to answer two research questions:

- RQ1: Can the addition of imperceptible noise/perturbation to logos allow these images to avoid detection by anti-phishing tools?
- RQ2: Which anti-phishing tools are better equipped to identify stolen logos when adversarial noise is introduced?

These questions are critical because phishing remains one of the most prevalent forms of cybercrime, with the potential to cause severe financial and data losses. According to Ji et al., "visual similarity-based detection models face challenges when adversarial samples are introduced" [1]. Phishing detection systems increasingly rely on visual similarity metrics to identify malicious logos; however, adversarial techniques like subtle perturbations in logos can bypass these systems. Addressing this vulnerability will improve the security of digital interactions and protect users from phishing attacks.

Our proposed approach involves generating perturbed logos using machine unlearning techniques, embedding these logos into websites, and systematically evaluating anti-phishing tools on their ability to detect both clean and noisy logos. This approach effectively addresses the weaknesses in visual similarity detection systems by testing their robustness against real-world adversarial attacks. As Shirazi et al. point out, "adversarial samples can successfully bypass trained phishing classifiers by altering key logo features" [2].

II. Methods: Data Collection

We will focus on tools that detect phishing based on visual similarity, particularly those that use logos to determine a site's authenticity. Participants in this study will be existing anti-phishing tools, such as Phish.AI and LogoGuard, rather than human subjects.

The anti-phishing tools will be exposed to both clean and noisy (perturbed) logos, and their detection accuracy will be measured. Specifically, we will:

- Generate perturbed logos from common brands such as Facebook, Instagram, and Bank of America using machine unlearning techniques.
- Create simple websites embedding these logos and subject them to testing by anti-phishing tools.

A. Data Collection

We will collect detection rates, false positive and false negative rates for both clean and perturbed logos.

B. Tools/Software

- Python: For generating perturbed logos.
- HTML/CSS: To create basic websites embedding the logos.
- Anti-phishing software: Tools such as Phish.AI and LogoGuard to evaluate detection effectiveness.

C. Validity

We will maintain consistency in factors unrelated to logos, such as URL structure and HTML content, to ensure that any differences in detection rates are due solely to the logo perturbations. This will ensure internal validity. We will test on a broad set of logos from different industries, increasing the external validity. Our metrics (detection rates, false positive/negative rates) are standard in phishing detection research, promoting construct validity.

D. Ethical Considerations

No human participants are involved, and no real-world phishing websites will be created. All experiments will be conducted in a secure, controlled environment.

III. Methods: Data Analysis

Our analysis will focus on both quantitative and qualitative methods. The study will adopt a quantitative approach, recording detection rates across various tools and conditions, with a focus on:

- Comparing the detection accuracy of anti-phishing tools for both clean and perturbed logos.

- Analyzing false positive and false negative rates.

We will also conduct qualitative analysis by examining specific logos that tools fail to detect and identifying why these perturbations bypass the detection systems.

A. Analysis Quality

We will use statistical methods to compare detection rates between clean and noisy logos, ensuring results are significant. Additionally, clean logos will serve as a baseline to ensure our comparisons are accurate.

IV. Anticipated Challenges and Limitations

One of the primary challenges we anticipate is the limited time available for the project, as we have less than two months to complete the entire study. Another challenge is our limited knowledge of machine learning and perturbation techniques, which may introduce delays in the data collection phase. To mitigate these risks, we will allocate sufficient time for learning and testing basic machine unlearning techniques early in the project timeline.

This study will focus on logo perturbations and will not cover other phishing techniques such as URL manipulation or malicious code insertion. However, as logos are a common target in phishing schemes, this focus is justified for our research scope.

V. Related Work

Prior work has explored phishing detection systems using visual similarity models. Ji et al. explored the weaknesses of phishing detection models under visual perturbations, concluding that "adversarial samples could evade many systems, lowering detection rates" [1]. Similarly, Hao et al. demonstrated that even slight visual distortions can significantly lower detection accuracy in logo-based phishing systems [3]. Charmet et al. examined how explanations could improve visual phishing detection, noting that "explanations are a crucial part of robust detection models" [4]. Our work builds on these studies by systematically testing existing tools against perturbed logos to identify specific improvements. Kulkarni et al. also noted that "robustness of phishing detection models can be compromised by adversarial attacks" [5], reinforcing the importance of our focus on perturbed logos.

VI. Budget and Timeline

A. Budget

The total budget for this project is zero, as all tools and resources will be open-source or free to use. We will not incur any costs for software, cloud resources, or data collection.

B. Team Roles

The overall team roles are distributed based on expertise:

- Reagan Sanz: As the team member with the most knowledge of machine unlearning, Reagan will take the lead in implementing the unlearning techniques and managing the technical aspects of the perturbation process.
- Faith Chernowski: With extensive experience in HTML development, Faith will lead the creation of the website used to test the logos. She will also manage website-related testing and presentation elements.
- Radhika Garg: Radhika will assist in data analysis and ensure that the results are well-documented. She will also handle organizing the timeline and final report preparation.

C. Timeline

Each team member is expected to dedicate approximately 4-6 hours per week to the project. Below is a detailed timeline:

Week 1 (Oct 1 - Oct 7):

- Finalize team roles, objectives, and scope of the project.
- Conduct a thorough literature review on machine unlearning, anti-phishing tools, and related work.
- Set up the project environment, including code repositories and initial dataset preparation.

Week 2 (Oct 8 - Oct 14):

- Gather logos from brands like Facebook, Instagram, and Bank of America for experimentation.
- Write the initial code to introduce imperceptible noise/perturbations to the logos using machine unlearning techniques.
- Start building a simple HTML/CSS website to showcase the clean and perturbed logos for testing.

Week 3 (Oct 15 - Oct 21):

- Perform preliminary testing using various anti-phishing tools (e.g., Phish.AI, LogoGuard) on noisy vs. clean logos.
- Collect initial results and begin documenting key findings.
- Prepare for the progress presentation on October 22nd, summarizing the research question, methodology, and preliminary results.

Week 4 (Oct 22 - Oct 28):

- Present progress on October 22nd, highlighting methodology, preliminary results, and next steps.
- Address any challenges encountered, such as noisy logos being detected by anti-phishing tools.

Week 5 (Oct 29 - Nov 4):

- Perform more in-depth testing and data analysis across different anti-phishing tools.
- Refine and expand on initial testing results, focusing on trends and detection rates.

Week 6 (Nov 5 - Nov 11):

- Finalize testing and ensure comprehensive data collection.
- Begin preparing the final presentation and start drafting the final report.

Week 7 (Nov 12 - Nov 18):

- Prepare the final presentation slide deck and refine visualizations, graphs, and key findings.

- Practice the final presentation and complete the final report.

November 21st:

- Deliver the final presentation and submit the completed report, highlighting key insights into anti-phishing tool robustness against perturbed logos.

VII. Conclusion

Our project aims to explore the vulnerability of anti-phishing tools against perturbed logos and provide valuable insights into improving detection robustness. We anticipate that our findings will highlight both the strengths and weaknesses of current tools, enabling more secure anti-phishing strategies in the future.

References

[1] F. Ji, K. Lee, H. Koo, W. You, E. Choo, H. Kim, and D. Kim, "Evaluating the effectiveness and robustness of visual similarity-based phishing detection models," *arXiv preprint arXiv:2405.19598v1*, 2024.

[2] H. Shirazi, B. Bezawada, I. Ray, and C. Anderson, "Adversarial sampling attacks against phishing detection," in *Proceedings of the 33rd IFIP Annual Conference on Data and Applications Security and Privacy*, 2019, pp. 83–101.

[3] Q. Hao, N. Diwan, Y. Yuan, G. Apruzzese, M. Conti, and G. Wang, "It doesn't look like anything to me: Using diffusion model to subvert visual phishing detectors," in *Proceedings of the 33rd USENIX Security Symposium*, 2024.

[4] F. Charmet, T. Morikawa, A. Tanaka, and T. Takahashi, "Vortex: Visual phishing detections are through explanations," *ACM Transactions on Internet Technology*, 2024.

[5] A. Kulkarni, V. Balachandran, D. M. Divakaran, and T. Das, "From ml to llm: Evaluating the robustness of phishing webpage detection models against adversarial attacks," *arXiv preprint arXiv:2407.20361v2*, 2024.