KRAKOW INTERNATIONAL CONFERENCE ON COMPUTER SCIENCE AND CYBERSECURITY

Detecting Phishing Campaigns Through Behavioral Analysis of Email Metadata

Piotr Zielinski
Faculty of Computer Science, AGH University of Krakow, Krakow, Poland
p.zielinski@example.org

Julia Novakova
Faculty of Informatics, Masaryk University, Brno, Czech Republic
j.novakova@example.org

DOI: 10.5281/zenodo.0000000

Abstract *— Phishing campaigns increasingly evade content-based detection filters by varying message text while preserving underlying sending infrastructure and behavioral patterns, motivating detection approaches that examine metadata rather than message content alone. This study develops a behavioral detection framework analyzing email metadata features, including sending time patterns, header inconsistencies, and relay path irregularities, to identify coordinated phishing campaigns across an organizational email corpus. Rather than classifying individual messages in isolation, the approach clusters messages exhibiting correlated metadata signatures to identify campaign-level patterns that individual message classifiers frequently miss, particularly for low-volume targeted campaigns designed to avoid triggering volume-based anomaly thresholds. We evaluate the framework against a labeled dataset combining confirmed phishing incidents with legitimate email traffic, comparing detection performance against a content-based baseline classifier under conditions simulating adversarial message text variation. Results show that the behavioral metadata approach maintained substantially more stable detection performance under text variation conditions where the content-based baseline degraded considerably, indicating greater robustness to evasion attempts targeting message content specifically. We discuss deployment considerations for integrating metadata-based campaign detection alongside existing content filtering within enterprise email security pipelines without introducing excessive false positive burden for security analysts.*

Index Terms *— phishing detection, email security, behavioral analysis, metadata analysis, cybersecurity*

I. INTRODUCTION

Phishing campaigns increasingly evade content-based detection filters by varying message text while preserving underlying sending infrastructure and behavioral patterns, motivating detection approaches that look beyond message content [1]. Content-based classifiers, though effective against unsophisticated attacks, degrade markedly when adversaries deliberately vary wording to avoid signature or classifier triggers [2].

Email metadata, including header structure, relay paths, and sending time patterns, offers signals that are comparatively costly for attackers to vary without disrupting operational infrastructure [3]. Low-volume targeted campaigns present a particular detection challenge, as they are designed specifically to avoid triggering volume-based anomaly thresholds common in enterprise defenses [4].

This study develops a behavioral detection framework that clusters messages by correlated metadata signatures to identify campaign-level phishing patterns, evaluated against a content-based baseline under adversarial text variation conditions.

Table 1. Detection performance under adversarial text variation

Condition	Metadata F1	Content-Based F1	False Positive
No variation	0.93	0.91	0.04
Low variation	0.91	0.78	0.05
Moderate variation	0.89	0.61	0.06
High variation	0.86	0.44	0.07

II. RELATED WORK

Content-based phishing detection has drawn extensively on natural language features and URL analysis, achieving strong performance against known attack patterns but exhibiting known brittleness to adversarial text perturbation. Metadata-based and infrastructure-based detection approaches have been explored in spam filtering more broadly, though their application specifically to campaign-level phishing clustering under text-variation adversarial conditions remains less thoroughly evaluated. Complementary evidence from adjacent literature further situates these dynamics within the broader field [5]. Related methodological precedents also inform the analytical choices adopted here [6].

III. METHODOLOGY

Email metadata features, including header inconsistencies, sending time distributions, and relay path irregularities, were extracted from an organizational email corpus combining confirmed phishing incidents with legitimate traffic.

A. Campaign-Level Clustering Approach

Rather than classifying individual messages in isolation, messages were clustered based on correlated metadata signatures using a similarity-based clustering algorithm, allowing identification of coordinated campaigns even when individual messages within a cluster exhibited highly varied textual content.

IV. RESULTS

Table 1 compares detection performance between the metadata-based clustering approach and a content-based baseline classifier across text variation conditions.

The metadata-based approach maintained substantially more stable detection performance as text variation intensity increased, while the content-based baseline degraded considerably.

V. DISCUSSION

The stability of metadata-based detection under adversarial text variation supports the interpretation that behavioral and infrastructural signals are more costly for attackers to obfuscate than message wording, offering a meaningful robustness advantage for campaign-level detection. The clustering approach additionally surfaced low-volume targeted campaigns that individual message classifiers frequently missed, addressing a known blind spot in volume-threshold-based anomaly detection. Practical deployment requires careful false positive management, since clustering-based flagging can implicate legitimate bulk communication patterns if not tuned to organizational context.

VI. CONCLUSION

Behavioral metadata analysis offers a robust complement to content-based phishing detection, particularly valuable against adversarial text variation and low-volume targeted campaigns. Enterprise email security pipelines should integrate metadata-based campaign detection alongside content filtering while managing false positive burden for security analysts.

REFERENCES

[1] S. Sheng et al., ”Who falls for phish? A demographic analysis of phishing susceptibility,” in Proc. CHI, 2010, pp. 373-382.

[2] R. Verma et al., ”Detecting phishing emails the natural language way,” in Proc. ESORICS, 2012, pp. 824-841.

[3] M. Chandrasekaran et al., ”Phishing email detection based on structural properties,” in Proc. NYS Cyber Secur. Conf., 2006.

[4] A. Almomani et al., ”A survey of phishing email filtering techniques,” IEEE Commun. Surv. Tutor., vol. 15, no. 4, pp. 2070-2090, 2013.

[5] G. Ho et al., ”Detecting and characterizing lateral phishing at scale,” in Proc. USENIX Secur., 2019, pp. 1273-1290.

[6] I. Fette et al., ”Learning to detect phishing emails,” in Proc. WWW, 2007, pp. 649-656.