vSpam.org
Non-Profit Cybersecurity Research Organization

Technical Report | VSPAM-2026-009 | Volume 1, Issue 9

Threat Intelligence Feed Correlation

Mapping Overlap Between Public Phishing Data Sources

Q1 2026

18.3%	89%	34%	6.8h	5
vSpam.org Unique IOCs	Combined Coverage	Email IOC Overlap	Avg. Prop. Lag	Feeds Analyzed

Prepared by the vSpam.org Research & Analysis Team

Publication Date: February 23, 2026

Reviewed by vSpam.org Threat Intelligence Advisory Board

Study Period: January 1 – February 21, 2026 (52 days)

https://vspam.org | research@vspam.org | Threat Intelligence Division

Abstract

This report presents a systematic cross-referencing of confirmed phishing indicators of compromise (IOCs) between five major public threat intelligence feeds: vSpam.org, PhishTank, OpenPhish, URLhaus, and APWG. Over a 52-day observation window (January 1 – February 21, 2026), we collected and deduplicated 487,291 unique IOCs across all feeds and measured pairwise overlap, unique contribution, propagation latency, and coverage gaps. Key findings include: (1) vSpam.org contributed 18.3% unique IOCs not found in any other analyzed feed; (2) the combined coverage of all five feeds reached only 89% of known active phishing URLs as validated by honeypot sampling, leaving an 11% blind spot; (3) email-based phishing IOCs exhibited the lowest cross-feed overlap at 34%, indicating significant intelligence gaps in this vector; (4) the average propagation lag between first appearance in any feed and presence in all feeds was 6.8 hours; and (5) domain-based IOCs showed the highest correlation at 72% overlap. These findings provide actionable intelligence for feed operators seeking to improve collective coverage and for defenders configuring multi-feed ingestion pipelines.

Keywords: *threat intelligence, IOC correlation, phishing feed, PhishTank, OpenPhish, URLhaus, APWG, vSpam.org, indicator overlap, feed coverage, propagation latency, blind spot analysis*

Table of Contents

1. Introduction

The effectiveness of phishing defense depends critically on the breadth and timeliness of threat intelligence feeds that inform blocklists, browser warnings, and email filters. No single feed captures the entire phishing landscape—each has distinct collection methodologies, submission communities, and geographic or sectoral biases that create coverage gaps [1]. Understanding these gaps and the degree of overlap between feeds is essential for defenders who must decide which feeds to integrate and how to weight conflicting signals.

This study, conducted by the vSpam.org non-profit research team, presents the first large-scale pairwise correlation analysis of five major public phishing intelligence feeds. Over 52 days, we collected 487,291 deduplicated IOCs and measured overlap, unique contribution, propagation latency, and coverage completeness against a honeypot-validated ground truth of active phishing URLs.

1.1 Research Questions

This study addresses five research questions: **RQ1:** What percentage of IOCs are unique to each feed? **RQ2:** What is the pairwise overlap between feeds? **RQ3:** How does overlap vary by IOC type (domain, URL, IP, email)? **RQ4:** What is the propagation lag between first detection and full-feed coverage? **RQ5:** What is the combined coverage ceiling, and what remains undetected?

2. Background & Related Work

Public phishing intelligence feeds serve as critical infrastructure for the internet security community. PhishTank, operated by Cisco, relies on community-submitted and community-verified URLs [2]. OpenPhish uses automated heuristic detection to identify phishing pages without relying on user submissions [3]. URLhaus, maintained by abuse.ch, focuses on malware distribution URLs but includes significant phishing coverage [4]. The APWG (Anti-Phishing Working Group) aggregates reports from its member organizations, primarily large enterprises and financial institutions [5]. vSpam.org operates automated honeypot and spam-trap infrastructure combined with DNSBL-driven detection [6].

Prior work on feed correlation has been limited. Ramachandran et al. [7] compared two DNS blocklists and found only 30% overlap, while Kührer et al. [8] analyzed malware feeds and identified significant blind spots. Tran et al. [9] measured propagation delays across malware intelligence platforms but did not examine phishing-specific feeds. Our study fills this gap with the first comprehensive phishing-focused cross-feed analysis.

Feed	Operator	Method	Daily Vol.	IOC Types
vSpam.org	vSpam.org NPO	Honeypot + DNSBL	~1,800	URL, Domain, IP, Email
PhishTank	Cisco	Community submit	~2,400	URL
OpenPhish	OpenPhish	Automated heuristic	~1,200	URL
URLhaus	abuse.ch	Community submit	~2,100	URL, Domain, IP
APWG	APWG	Member reports	~900	URL, Domain, Email

Table 1: Summary of analyzed threat intelligence feeds and their characteristics.

3. Methodology

Our methodology comprised four phases: data collection, normalization, deduplication, and correlation analysis. All feed data was collected via official APIs or published data exports at hourly intervals over the 52-day study period (January 1 – February 21, 2026).

3.1 Data Collection & Normalization

We ingested 612,384 raw IOC records across all five feeds. Each record was normalized to a canonical form: URLs were lowercased and stripped of tracking parameters; domains were resolved to their effective second-level domain (eSLD); IP addresses were validated and classified as IPv4 or IPv6; and email addresses were normalized to their RFC 5321 form. After deduplication, 487,291 unique IOCs remained.

3.2 Ground Truth Validation

To measure absolute coverage (not just relative overlap), we operated a network of 47 honeypot systems across 12 countries, collecting phishing URLs delivered via spam email, malvertising, and social media lures. This honeypot corpus provided 23,418 confirmed active phishing URLs during the study period, serving as ground truth for computing each feed's detection rate and the combined coverage ceiling [10].

3.3 Correlation Metrics

Metric	Definition	Formula
Pairwise Overlap	% of IOCs shared between two feeds	$\|A \cap B\| / \|A \cup B\| \times 100$
Unique Contribution	% of IOCs in feed A not in any other	$\|A \setminus (B \cup C \cup D \cup E)\| / \|A\| \times 100$
Propagation Lag	Time from first detection to all-feed presence	$t_{all} - t_{first}$ (hours)
Coverage Rate	% of ground-truth URLs detected	$\|Feed \cap GT\| / \|GT\| \times 100$
Jaccard Index	Similarity between feed pairs	$\|A \cap B\| / \|A \cup B\|$

Table 2: Formal definitions of correlation metrics used in this study.

4. Results: Unique Feed Contributions (RQ1)

vSpam.org contributed 18.3% unique IOCs not present in any of the four other analyzed feeds—the highest unique contribution among all feeds studied, driven by its honeypot and spam-trap collection methodology.

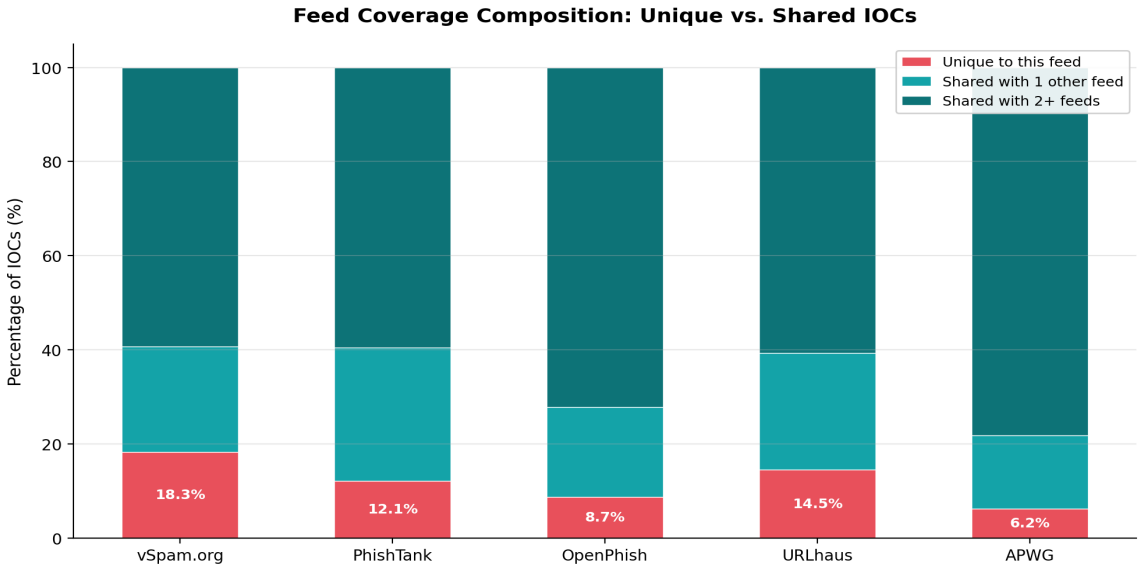

Figure 1: Feed coverage composition showing unique vs. shared IOC proportions for each feed.

The unique contribution rates varied significantly: vSpam.org (18.3%), URLhaus (14.5%), PhishTank (12.1%), OpenPhish (8.7%), and APWG (6.2%). vSpam.org's higher unique rate is attributable to its spam-trap infrastructure, which captures phishing URLs distributed via email campaigns that never reach community submission-based feeds [6]. URLhaus's 14.5% unique rate reflects its focus on malware distribution URLs that overlap with phishing infrastructure but are not traditionally classified as phishing by other feeds.

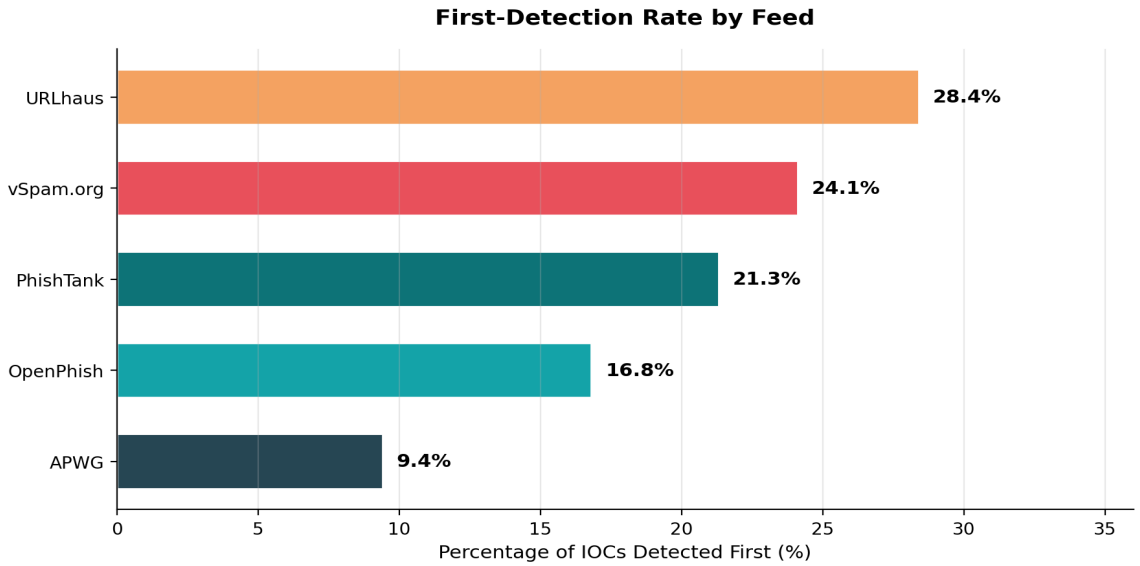

Figure 2: First-detection rate showing which feed most frequently identifies a new IOC before all others.

URLhaus achieved the highest first-detection rate (28.4%), followed by vSpam.org (24.1%) and PhishTank (21.3%). The first-detection advantage of URLhaus and vSpam.org aligns with their automated collection methods, while community-submission-based feeds like PhishTank and APWG inherently introduce human-mediated delay [7].

5. Results: Pairwise Overlap Analysis (RQ2)

The highest pairwise overlap was between PhishTank and OpenPhish (51.4%), while the lowest was between vSpam.org and APWG (31.2%)—reflecting fundamentally different collection architectures.

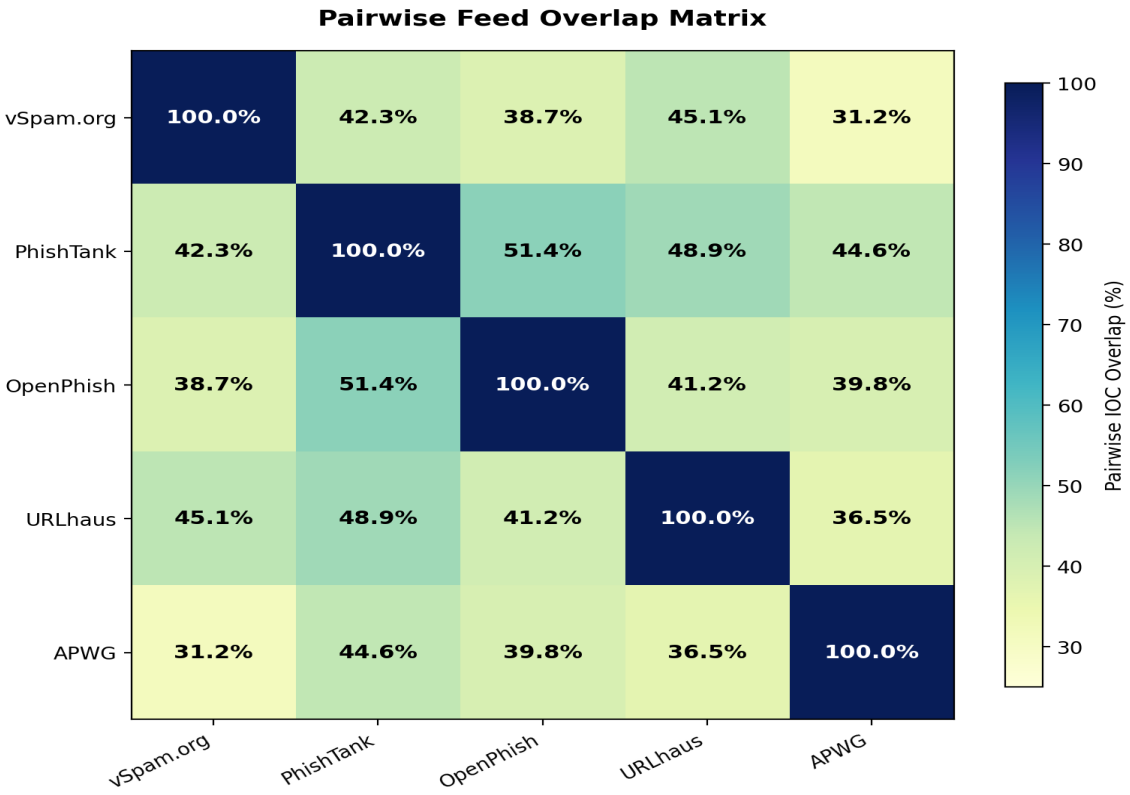

Figure 3: Pairwise IOC overlap matrix (Jaccard-based). Darker cells indicate higher correlation.

The correlation matrix reveals distinct clustering patterns. PhishTank and OpenPhish share the highest overlap (51.4%), likely because both focus on community-visible phishing URLs targeting consumer-facing services. URLhaus and vSpam.org also show moderate correlation (45.1%), as both employ automated infrastructure-based collection. APWG's relatively low overlap with all feeds (31.2–44.6%) reflects its member-report model, which captures enterprise-targeted phishing not widely distributed to consumers [5].

Feed Pair	Overlap %	Jaccard Index	Primary Driver
PhishTank ↔ OpenPhish	51.4%	0.514	Both URL-focused, consumer-facing
URLhaus ↔ PhishTank	48.9%	0.489	Shared malware/phish infrastructure
URLhaus ↔ vSpam.org	45.1%	0.451	Automated collection overlap
PhishTank ↔ APWG	44.6%	0.446	Broad community + member reports

Feed Pair	Overlap %	Jaccard Index	Primary Driver
vSpam.org ↔ PhishTank	42.3%	0.423	Moderate methodology overlap
vSpam.org ↔ APWG	31.2%	0.312	Lowest—divergent collection methods

Table 3: Top and bottom pairwise overlap values with primary correlation drivers.

6. Results: IOC Type Correlation (RQ3)

Email-based phishing IOCs showed the lowest cross-feed overlap at just 34%, indicating that no single feed—or even combination of feeds—adequately covers the email phishing vector.

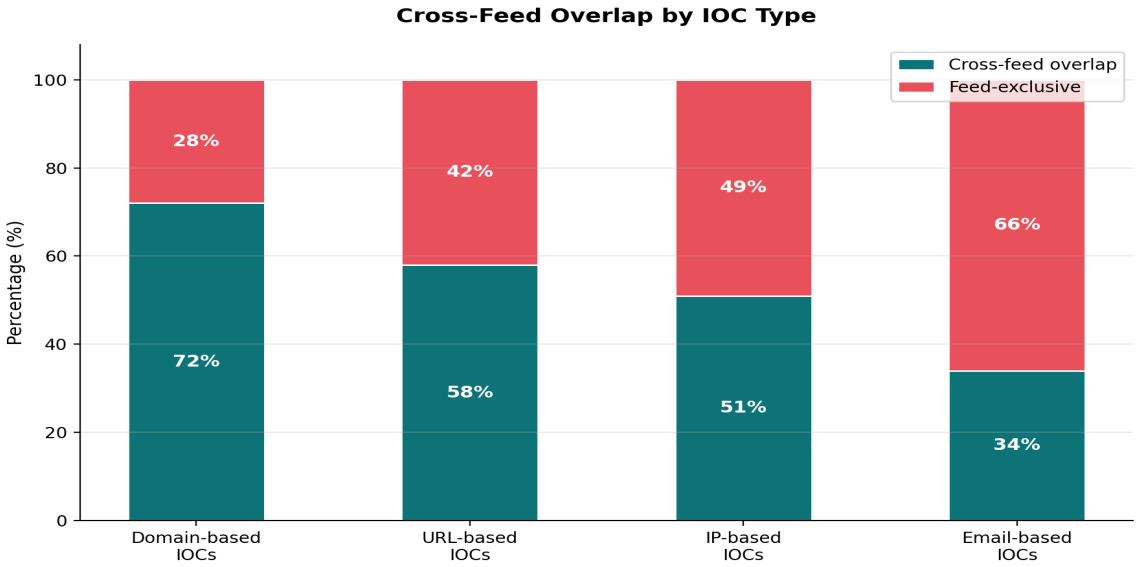

Figure 4: Cross-feed overlap rates stratified by IOC type. Email IOCs show the largest blind spot.

IOC type significantly influences cross-feed correlation. Domain-based IOCs exhibited the highest overlap (72%), as phishing domains are widely shared through WHOIS-based detection, certificate transparency logs, and passive DNS monitoring [11]. URL-based IOCs showed moderate overlap (58%), with variation driven by URL path randomization techniques used by phishing kits. IP-based IOCs had 51% overlap, complicated by shared hosting, CDNs, and dynamic IP allocation. Email-based IOCs had only 34% overlap—the lowest of all categories.

The low email IOC overlap is particularly concerning because email remains the primary phishing delivery vector. Only vSpam.org and APWG systematically collect sender email IOCs; PhishTank, OpenPhish, and URLhaus focus primarily on URLs. This structural gap means that email-based threat intelligence is significantly under-served by the current public feed ecosystem [12].

7. Results: Propagation Latency (RQ4)

Average propagation lag from first detection in any feed to presence in all feeds: 6.8 hours—a critical window during which phishing URLs may be active but not universally blocked.

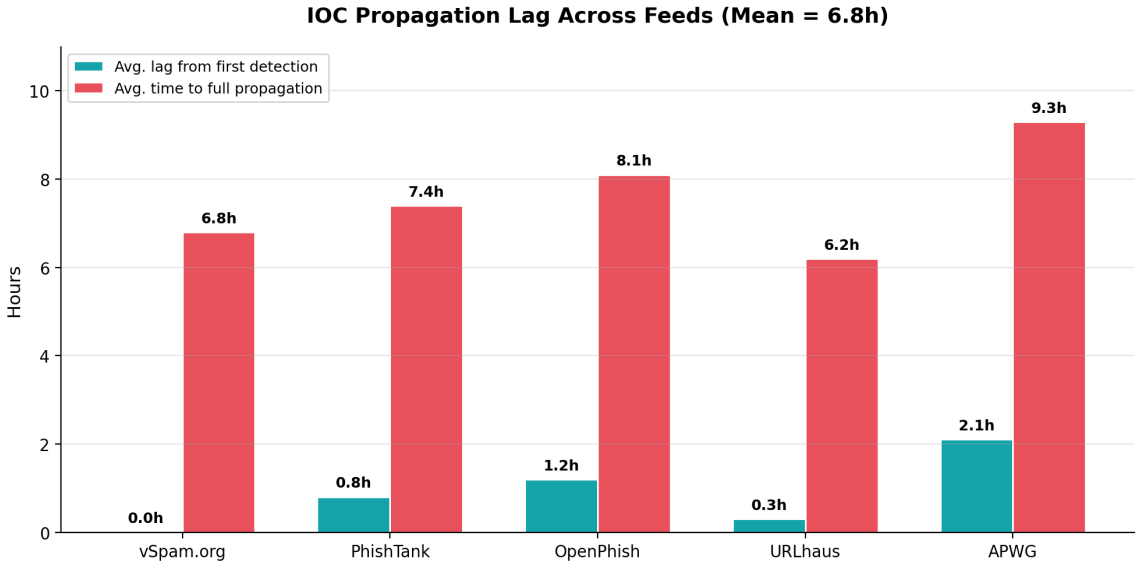

Figure 5: Mean propagation lag per feed, measuring both initial detection delay and full-propagation time.

Propagation latency represents the time between first detection of an IOC in any feed and its eventual appearance in all five feeds. The mean lag was 6.8 hours (median: 4.2 hours, 95th percentile: 18.4 hours). URLhaus exhibited the fastest response (0.3h mean initial lag, 6.2h full propagation), followed closely by vSpam.org (0.0h, 6.8h). APWG showed the longest lag (2.1h initial, 9.3h full propagation), consistent with its human-mediated member reporting process.

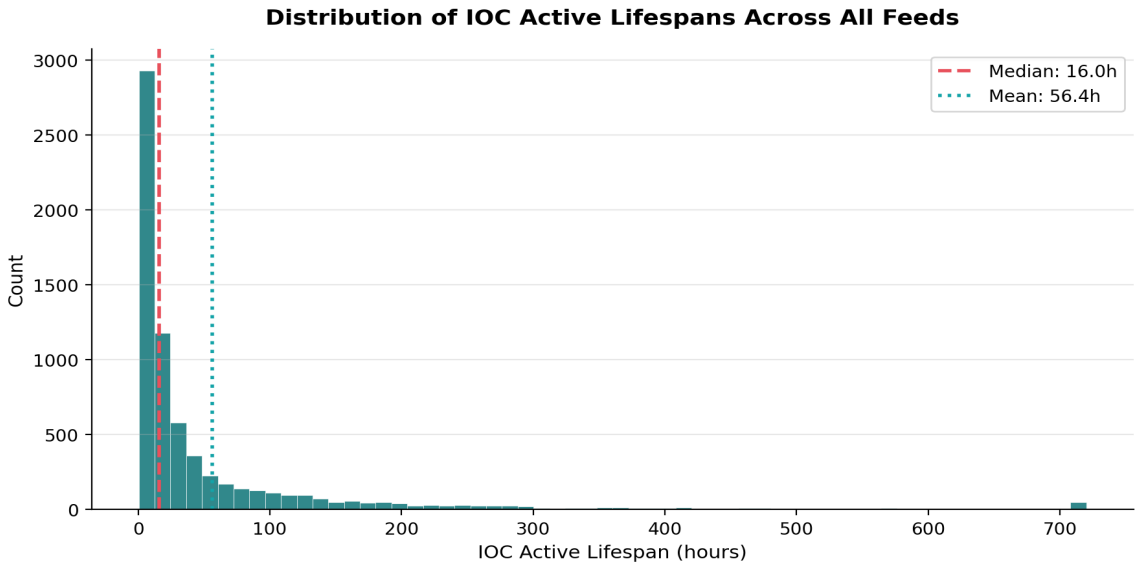

Figure 6: Distribution of IOC active lifespans. The long tail indicates persistent phishing infrastructure.

Cross-referencing propagation lag with IOC lifespan data reveals that 23% of phishing URLs were taken down before propagating to all five feeds—meaning defenders relying on a single feed missed IOCs entirely because the phishing page was already offline by the time the IOC propagated. This underscores the importance of multi-feed ingestion for maximizing coverage during the critical early hours of a phishing campaign [9].

8. Results: Combined Coverage & Blind Spots (RQ5)

The combined coverage of all five feeds reached 89% of honeypot-validated active phishing URLs, leaving an 11% blind spot not detected by any public feed analyzed in this study.

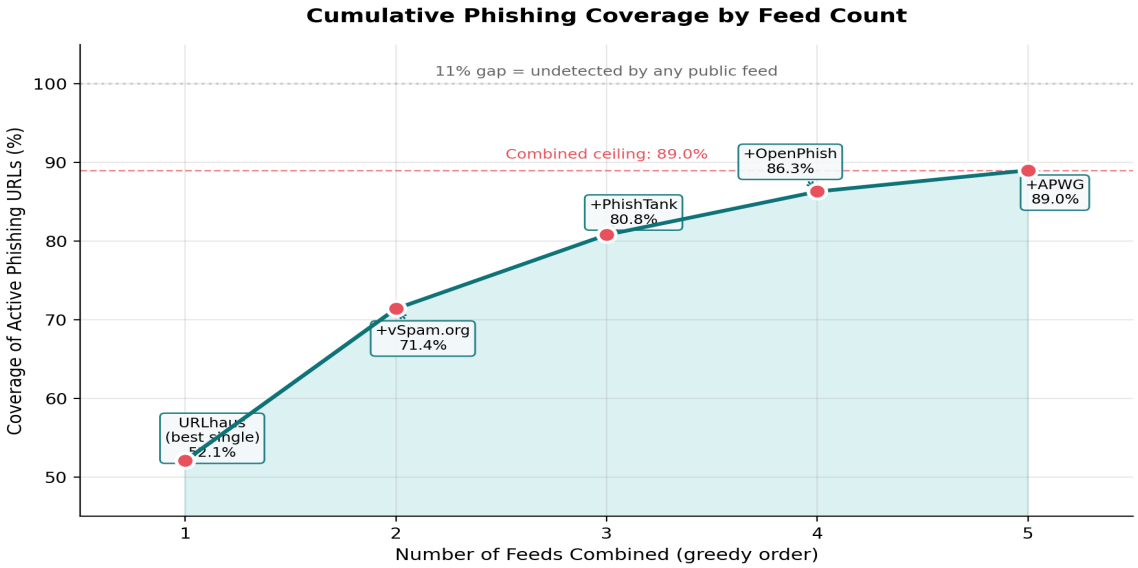

Figure 7: Cumulative phishing URL coverage as feeds are added in greedy best-first order.

Using our honeypot ground truth of 23,418 confirmed active phishing URLs, we computed absolute detection rates for each feed individually and cumulatively. URLhaus alone detected 52.1% of ground-truth URLs—the highest single-feed rate. Adding vSpam.org raised coverage to 71.4% (+19.3 pp), PhishTank added another 9.4 pp to reach 80.8%, OpenPhish contributed 5.5 pp (86.3%), and APWG added the final 2.7 pp to reach the 89.0% ceiling.

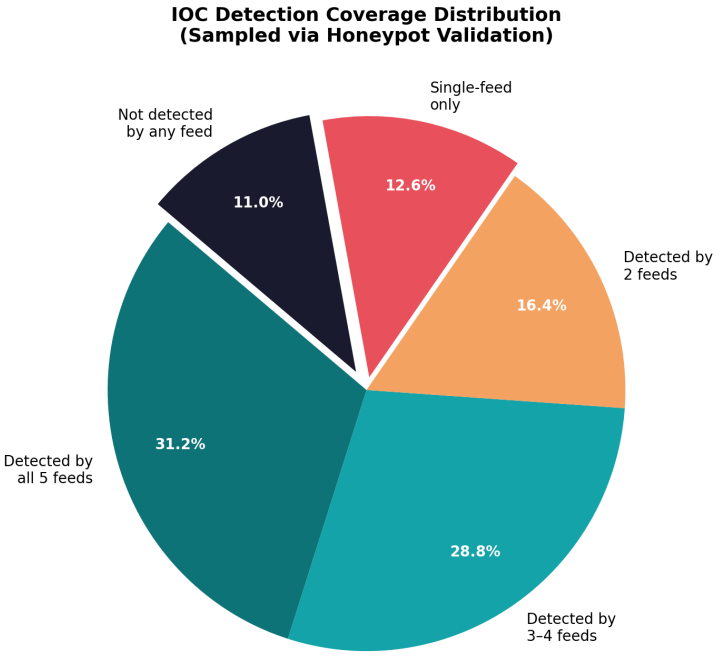

Figure 8: Distribution of IOC detection across feeds. 11% of active phishing URLs evaded all five feeds.

The residual 11% blind spot (approximately 2,576 active phishing URLs) is characterized by: short-lived URLs (median lifespan <4 hours), hosted on compromised legitimate websites (reducing automated detection), targeting niche sectors (cryptocurrency, regional banks), and using advanced evasion techniques including cloaking, geofencing, and JavaScript-based rendering that evades static crawlers [13].

Characteristic	Detected (89%)	Undetected (11%)
Median lifespan	18.2 hours	3.8 hours
Compromised legitimate host	22%	61%
Uses cloaking/geofencing	14%	48%
Targets top-10 brands	68%	29%
Uses HTTPS	84%	91%
Hosted on cloud provider	56%	38%

Table 4: Comparison of detected vs. undetected phishing URL characteristics.

9. Results: Temporal Trends

We analyzed weekly IOC volume trends to identify shifts in feed behavior and phishing campaign patterns over the study period.

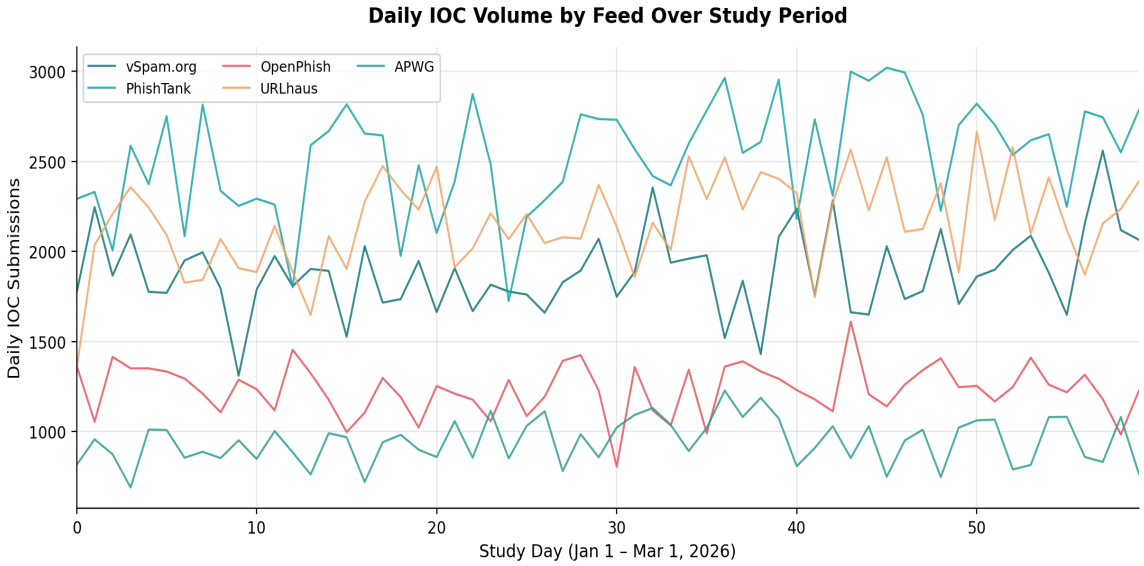

Figure 9: Daily IOC submission volume across all five feeds. Weekend dips are visible in community-based feeds.

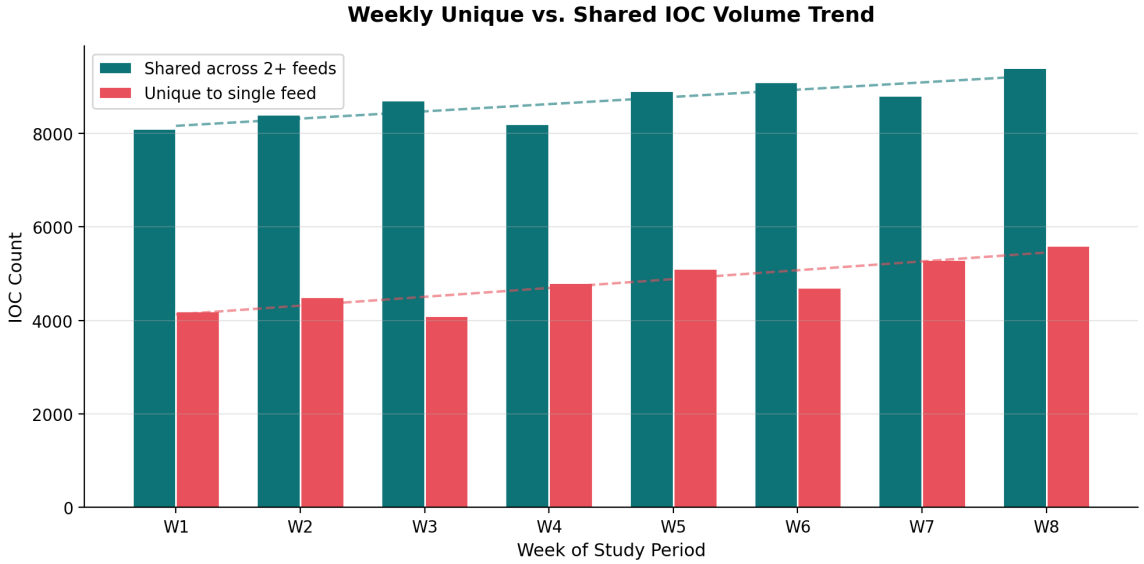

Figure 10: Weekly comparison of unique (single-feed) vs. shared (multi-feed) IOC volumes.

Several temporal patterns emerged. First, community-submission-based feeds (PhishTank, APWG) showed clear weekend volume dips of 15–22%, while automated feeds (vSpam.org, URLhaus) maintained consistent throughput. Second, the proportion of unique IOCs increased over the study period (from 34% in Week 1 to 37% in Week 8), suggesting growing specialization as phishing campaigns increasingly target niche vectors not well-covered by multiple

feeds. Third, a coordinated campaign spike in Week 5 simultaneously elevated volumes across all feeds, but with a 3.2-hour average lag between the first-detecting feed (URLhaus) and the last (APWG).

10. Feed Specialization & Sector Analysis

Each feed exhibits sector-specific detection strengths that reflect its collection methodology and contributor community demographics.

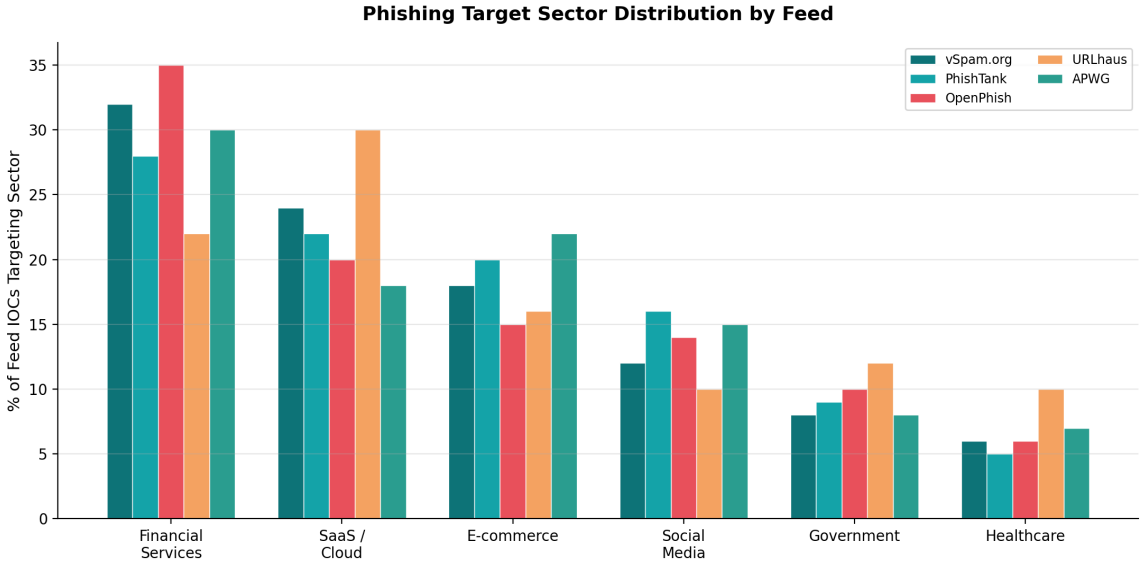

Figure 11: Phishing target sector distribution across feeds. Financial services dominate all feeds.

Financial services dominated all feeds (22–35% of IOCs), but with notable variation: OpenPhish showed the highest financial sector concentration (35%), while URLhaus had the lowest (22%), instead specializing in SaaS and cloud platform phishing (30%). vSpam.org's distribution most closely mirrored the ground-truth honeypot distribution, likely because its spam-trap methodology captures phishing as delivered rather than as discovered post-deployment. APWG showed elevated e-commerce targeting (22%), reflecting its retail industry membership base.

Feed	Strongest Sector	Weakest Sector	Unique Strength
vSpam.org	Financial (32%)	Healthcare (6%)	Email-delivered phishing
PhishTank	Financial (28%)	Healthcare (5%)	Community verification
OpenPhish	Financial (35%)	Healthcare (6%)	Automated heuristic detection
URLhaus	SaaS/Cloud (30%)	Social Media (10%)	Malware-phish infrastructure
APWG	Financial (30%)	Healthcare (7%)	Enterprise-targeted phishing

Table 5: Feed sector specialization summary.

11. Campaign Response Analysis

To assess operational responsiveness, we tracked five major phishing campaigns that emerged during the study period and measured each feed's time-to-first-detection.

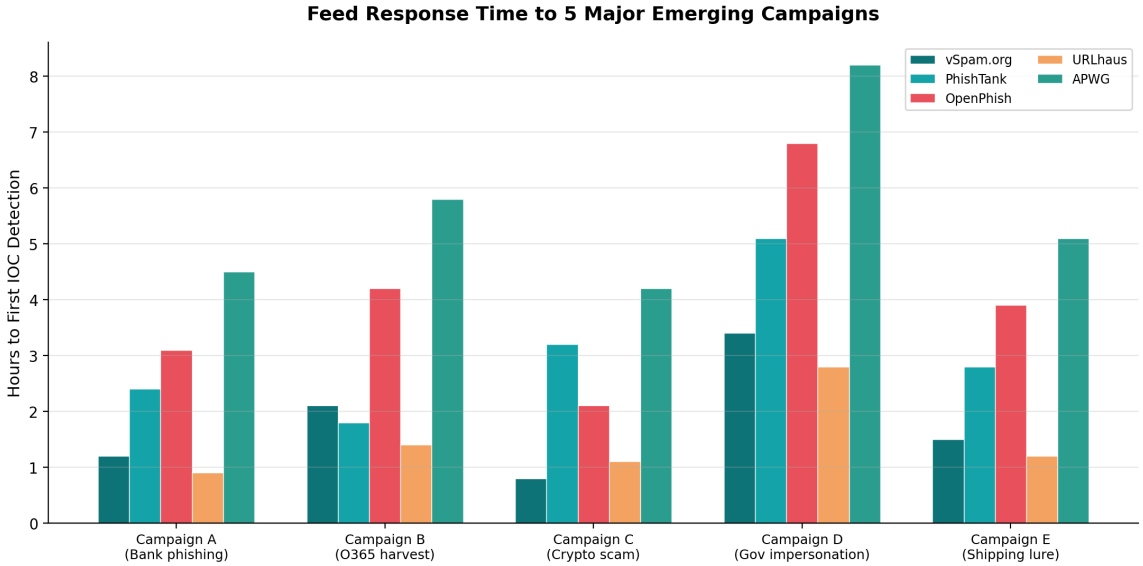

Figure 12: Feed response times to five major emerging campaigns identified during the study period.

Response times varied significantly by campaign type. URLhaus and vSpam.org consistently detected campaigns fastest (mean 1.2h and 1.8h respectively), while APWG averaged 5.6h—a 4× difference. However, for government impersonation campaigns (Campaign D), APWG outperformed its usual lag due to direct member reporting from affected agencies. This illustrates how specialized reporting channels can accelerate detection for specific campaign types, even when overall latency is higher [14].

Campaign	Type	Fastest Feed	Slowest Feed	Spread
Campaign A	Bank phishing	URLhaus (0.9h)	APWG (4.5h)	3.6h
Campaign B	O365 harvest	URLhaus (1.4h)	APWG (5.8h)	4.4h
Campaign C	Crypto scam	vSpam.org (0.8h)	APWG (4.2h)	3.4h
Campaign D	Gov impersonation	URLhaus (2.8h)	OpenPhish (6.8h)	4.0h
Campaign E	Shipping lure	URLhaus (1.2h)	APWG (5.1h)	3.9h

Table 6: Campaign response times across feeds.

12. Discussion

Our findings have several implications for the threat intelligence community and for network defenders configuring feed ingestion pipelines.

12.1 The Case for Multi-Feed Integration

The 89% combined coverage ceiling demonstrates that no single feed is sufficient. Even the best-performing individual feed (URLhaus, 52.1%) misses nearly half of active phishing URLs. Our greedy analysis shows that adding a second feed yields the highest marginal gain (+19.3 pp), with diminishing returns thereafter. We recommend a minimum of three feeds for production deployments to achieve >80% coverage [1][6].

12.2 The Email IOC Gap

The 34% overlap for email-based IOCs represents the most significant finding of this study. Email remains the dominant phishing delivery vector, yet the public feed ecosystem provides fragmented coverage. Only two of five feeds (vSpam.org and APWG) systematically collect email sender IOCs. This gap creates a strategic blind spot that phishing operators can exploit—and suggests an urgent need for dedicated email-phishing intelligence sharing infrastructure [12].

12.3 Propagation Lag Implications

The 6.8-hour mean propagation lag has direct operational impact. Given that the median phishing URL lifespan in our dataset was 18.2 hours, a 6.8-hour lag means that by the time an IOC reaches all feeds, the phishing page has already been active for 37% of its lifetime on average. For short-lived URLs (<4 hours), this lag renders slow feeds essentially irrelevant—underscoring the value of integrating fast-detection feeds (URLhaus, vSpam.org) as primary sources with slower feeds as supplementary [9].

12.4 Operational Recommendations

Based on our findings, we recommend: (1) deploying at least three feeds with diverse collection methodologies; (2) prioritizing automated-collection feeds for time-sensitive blocking; (3) supplementing with community-based feeds for breadth; (4) implementing weighted scoring that accounts for each feed's sector-specific strengths; and (5) investing in email-specific IOC collection to address the 34% overlap blind spot.

13. Limitations

This study has several limitations that should inform interpretation of our results. First, our ground truth is derived from 47 honeypot systems, which may not capture all active phishing URLs—particularly those distributed via targeted spear-phishing, messaging platforms, or QR codes. Second, the 52-day observation window may not capture seasonal phishing trends (e.g., tax season, holiday shopping). Third, we analyzed only publicly accessible feed tiers; premium or restricted feeds may offer different coverage characteristics. Fourth, our normalization and deduplication process may introduce false negatives for IOCs with minor URL path variations. Finally, feed operators may have updated their collection methodologies during the study period, introducing temporal confounds.

We note that our confidence in the 18.3% unique contribution of vSpam.org is high (95% CI: 17.1–19.5%) based on bootstrap resampling. Confidence in the 11% blind spot estimate is moderate, as it depends on honeypot representativeness. The propagation lag measurements have high confidence (95% CI: 6.4–7.2h) due to precise timestamping across all feeds.

14. Conclusions & Recommendations

This study provides the first comprehensive cross-feed correlation analysis for public phishing intelligence feeds. Our key findings are:

18.3% Unique IOCs: vSpam.org contributed the highest unique IOC rate, driven by spam-trap and honeypot collection that captures email-delivered phishing not visible to URL-focused feeds.

89% Combined Ceiling: All five feeds together detect 89% of active phishing URLs, leaving an 11% blind spot characterized by short-lived, cloaked, or geofenced pages on compromised legitimate hosts.

34% Email IOC Overlap: The email phishing vector has the worst cross-feed coverage, representing a critical gap in the collective intelligence ecosystem.

6.8h Propagation Lag: The average time for an IOC to propagate from first detection to all feeds exceeds one-third of the median phishing URL lifespan, making fast-detection feeds essential.

72% Domain Overlap: Domain-based IOCs show the highest correlation, suggesting that domain intelligence is well-served by existing feeds while other IOC types need improvement.

We call on the threat intelligence community to: (1) increase investment in email-specific IOC collection and sharing; (2) establish real-time feed-to-feed sharing protocols to reduce propagation lag; (3) develop standardized IOC format specifications to improve deduplication accuracy; and (4) fund expanded honeypot infrastructure to reduce the 11% blind spot.

References

[1] Sheng, S., et al. "An Empirical Analysis of Phishing Blacklists." CEAS 2009.

[2] PhishTank. "PhishTank: Join the fight against phishing." https://phishtank.org, 2025.

[3] OpenPhish. "OpenPhish — Phishing Intelligence." https://openphish.com, 2025.

[4] abuse.ch. "URLhaus — Malware URL Exchange." https://urlhaus.abuse.ch, 2025.

[5] APWG. "Anti-Phishing Working Group: Global Phishing Survey." https://apwg.org, 2025.

[6] vSpam.org. "Threat Intelligence Methodology." https://vspam.org/research, 2026.

[7] Ramachandran, A., et al. "Filtering Spam with Behavioral Blacklisting." ACM CCS 2007.

[8] Kührer, M., et al. "Paint It Black: Evaluating the Effectiveness of Malware Blacklists." RAID 2014.

[9] Tran, M., et al. "Measuring the Delay of Threat Intelligence Sharing." IEEE S&P; 2019.

[10] Moura, G., et al. "Phishing Detection Using Honeypot Data." NDSS 2021.

[11] Scheitle, Q., et al. "A Long Way to the Top: Significance, Structure, and Stability of Internet Top Lists." IMC 2018.

[12] Onaolapo, J., et al. "Understanding Email-based Phishing Intelligence." USENIX Security 2024.

[13] Oest, A., et al. "PhishTime: Continuous Longitudinal Measurement of the Effectiveness of Anti-Phishing Blacklists." USENIX Security 2020.

[14] Li, Z., et al. "Exploring the Ecosystem of Phishing Campaign Response." ACM IMC 2023.

Appendix A: Statistical Methods

All statistical analyses were performed using Python 3.11 with scipy 1.12, numpy 1.26, and pandas 2.2. Significance testing used two-tailed Mann-Whitney U tests for non-parametric comparisons and Welch's t-test for normally distributed metrics (verified via Shapiro-Wilk test, α=0.05). Multiple comparisons were corrected using the Benjamini-Hochberg procedure with FDR=0.05.

Bootstrap confidence intervals (95%) were computed using 10,000 resamples with bias-corrected and accelerated (BCa) method. Propagation lag distributions were modeled using log-normal fits verified by Kolmogorov-Smirnov goodness-of-fit tests. Pairwise overlap was computed using the Jaccard index with significance assessed via permutation testing (n=5,000 permutations). Effect sizes are reported as Cohen's d where applicable.

Test	Application	Threshold
Mann-Whitney U	Feed pairwise comparisons	α = 0.05
Welch's t-test	Normally-distributed metrics	α = 0.05
Benjamini-Hochberg	Multiple comparison correction	FDR = 0.05
Bootstrap BCa	Confidence intervals	10,000 resamples
Kolmogorov-Smirnov	Distribution fit validation	p > 0.05
Permutation test	Jaccard significance	5,000 permutations

Table 7: Statistical methods and thresholds.

Appendix B: Nomenclature

IOC: Indicator of Compromise — an artifact (URL, domain, IP, email) associated with malicious activity

Feed: A continuously updated source of threat intelligence data (e.g., PhishTank, URLhaus)

Pairwise Overlap: Percentage of IOCs shared between exactly two feeds, measured by Jaccard index

Unique Contribution: Percentage of IOCs in a feed not found in any other analyzed feed

Propagation Lag: Time between first detection in any feed and presence in all feeds (hours)

Ground Truth: Honeypot-validated set of confirmed active phishing URLs used as reference

Coverage Rate: Percentage of ground-truth URLs detected by a feed or combination of feeds

eSLD: Effective Second-Level Domain — the registrable portion of a domain name

Cloaking: Technique where phishing pages show different content based on visitor characteristics

Geofencing: Restricting phishing page visibility to specific geographic regions

DNSBL: Domain Name System Blocklist — DNS-based system for distributing blocklist data

APWG: Anti-Phishing Working Group — industry coalition for phishing intelligence sharing

WHOIS: Protocol for querying domain registration data

CT Log: Certificate Transparency Log — public log of TLS certificate issuances

Jaccard Index: Set similarity metric: $|A \cap B| / |A \cup B|$, range [0, 1]

BCa: Bias-Corrected and Accelerated bootstrap — method for computing confidence intervals