DOM-SP-SHAP: Structure-Preserving Explainability for Graph-Based Phishing Detection

Carlos Güemes-Palau*†, Pedro Casas†, Tailai Song†‡, Lucas Torrealba†§, Michela Meo‡ and Pere Barlet-Ros*

†*AIT Austrian Institute of Technology*, Vienna, Austria
**Barcelona Neural Networking Center, Universitat Politècnica de Catalunya*, Barcelona, Spain
‡*Politecnico di Torino*, Turin, Italy
§*NIC Chile Research Labs*, Santiago, Chile
{carlos.guemes,pere.barlet}@upc.edu, pedro.casas@ait.ac.at, lucas@niclabs.cl, {tailai.song,michela.meo}@polito.it

***Abstract*—Graph-based phishing detectors have recently shown strong performance by modeling the structural properties of webpages through Document Object Model (DOM) representations. However, these models remain largely opaque, limiting interpretability, analyst trust, and forensic understanding of their predictions. We present DOM-SP-SHAP, a structure-preserving explainability framework for graph-based phishing detection. DOM-SP-SHAP extends SHAP-based attribution to hierarchical DOM graphs while preserving the integrity of the underlying tree structure during perturbation. The proposed approach enables fine-grained node-level explanations for deep graph models without violating parent-child relationships or disrupting structural semantics. We evaluate DOM-SP-SHAP on SpecularNet, a GNN-based model for reference-free phishing detection, across three open phishing datasets spanning different years, including both real-world and adversarial webpages. Our analysis reveals highly consistent attribution patterns across datasets: phishing detection is driven primarily by shallow DOM regions associated with global webpage layout and structural organization, while deeper content-level nodes contribute minimally to predictions. These findings show that graph-based phishing detectors rely on stable structural invariants rather than superficial textual or visual cues, helping explain the strong detection performance achieved by DOM graph-based models such as SpecularNet.**

I. Introduction

Phishing remains one of the most pervasive cyber threats on the Web, motivating the development of increasingly sophisticated detection systems. Traditional defenses such as heuristic methods and blocklists are lightweight and interpretable but largely reactive, often failing against rapidly evolving phishing campaigns. More recent machine learning and deep learning approaches improve detection accuracy by learning patterns from URLs and webpage content, while reference-based systems leverage external brand knowledge and visual similarity to achieve state-of-the-art performance. However, these approaches often suffer from high computational overhead, limited deployability, and poor interpretability.

To address these limitations, reference-free structural phishing detection has emerged as a promising alternative. By modeling intrinsic webpage properties such as Document Object Model (DOM) structure and domain-name patterns, these approaches capture structural invariants of phishing kits that are more stable and difficult to manipulate than superficial textual or visual content. In this context, we recently introduced SpecularNet [1], [2] a hierarchical graph autoencoding approach that models webpages as DOM trees and achieves strong phishing-detection performance without relying on external references.

Despite its effectiveness, SpecularNet remains largely opaque. As a deep graph-based model relying on reconstruction-driven reasoning, it is difficult to identify which structural components contribute most to its predictions. This lack of interpretability limits analyst trust, hinders forensic validation, and obscures potential adversarial weaknesses. Existing explainability techniques such as LIME [3] and SHAP [4] have been applied to phishing detection and graph neural networks through methods such as GNNExplainer [5], PGM-Explainer [6], and SubgraphX [7]. However, these approaches are not specifically designed for hierarchical DOM trees and often rely on perturbations that break structural validity.

We introduce *DOM-SP-SHAP*, a structure-preserving explainability framework for graph-based phishing detection. DOM-SP-SHAP extends SHAP-based attribution to hierarchical DOM graphs while preserving the integrity of the underlying tree structure during perturbation. The proposed approach enables fine-grained node-level explanations for graph autoencoder architectures without violating parent–child relationships or disrupting structural semantics. Using DOM-SP-SHAP, we analyze the behavior of SpecularNet and show that phishing detection is primarily driven by shallow DOM regions associated with webpage layout and structural organization.

II. SpecularNet and DOM-SP-SHAP

A. SpecularNet: Hierarchical Graph Autoencoding

SpecularNet models a webpage's HTML DOM as a tree-structured graph and learns structural invariants through a hierarchical graph autoencoder [1]. The model relies exclusively on intrinsic webpage signals: the DOM structure and the domain name. At the input, the domain name is encoded at the character level using embeddings aggregated through an LSTM, while the HTML is represented as a tree of tags and attributes, discarding textual content to improve generalization.

978-3-903176-83-6 ©2026 IFIP

Algorithm 1 DOM Structure-Preserving SHAP

Require: $\mathcal{G} = (\mathcal{V}, \mathcal{E}), \text{Model}, v_0, k$
Ensure: $s_i, \forall v_i \in \mathcal{V}$
$s'_i \leftarrow 0, \forall v_i \in \mathcal{V}$
for $t = 1$ to k **do**
  $perm \leftarrow \text{RandomPermutation}(\mathcal{V} \setminus \{v_0\})$
  $\mathcal{G}' \leftarrow \mathcal{G}$
  $o \leftarrow \text{Model}(\mathcal{G})$
  for all $v_i \in perm$ **do**
    reconnect children of v_i to closest valid ancestor
    remove v_i from $\mathcal{G}'$
    $o' \leftarrow \text{Model}(\mathcal{G}')$
    $s'_i \leftarrow s'_i + (o - o')$
    $o \leftarrow o'$
  end for
end for
$s_i \leftarrow s'_i / k$

The DOM graph is processed through Graph ConvNet layers followed by root-preserving top-k pooling operations to retain the most informative structural components.

The core of SpecularNet consists of a hierarchical graph autoencoder that performs bottom-up message passing to compress structural information into a root-level representation, followed by top-down reconstruction of the DOM tree. Detection combines reconstruction-based anomaly scoring with supervised classification: benign webpages tend to reconstruct accurately, whereas phishing pages exhibit higher structural distortion. An important consideration for interpreting the explainability results presented next is that SpecularNet uses inverted labels, where higher output probabilities (> 0.5) correspond to benign webpages and lower probabilities (< 0.5) to phishing pages. Consequently, positive SHAP values indicate benign behavior, while negative values push predictions toward the phishing class. This formulation naturally follows from the model's reconstruction-based design, which is optimized to reconstruct benign webpage structures.

B. DOM-Structure-Preserving SHAP

Applying SHAP to hierarchical DOM trees is non-trivial, as naive perturbations can break the tree structure and invalidate the input graph. DOM-SP-SHAP is a structure-preserving extension of SHAP tailored to DOM-based graph models. This model builds upon the SHAP framework [4], which estimates feature contributions by measuring the impact of perturbations on the model output. While SHAP is model-agnostic and theoretically grounded, its standard formulation assumes independent features and does not account for hierarchical dependencies inherent to DOM trees.

DOM-SP-SHAP overcomes this limitation by redefining perturbations as structure-preserving transformations of the DOM graph. Specifically, nodes are removed without breaking the tree structure: when a non-leaf node is deleted, its children are deterministically reconnected to the closest valid ancestor in the original tree. This preserves parent-child relationships and maintains structural consistency throughout the attribution process. The procedure is also aligned with SpecularNet's inference mechanism, which similarly preserves the root node during graph pooling operations.

Algorithm 1 summarizes the proposed approach. Given a DOM graph $\mathcal{G}$, DOM-SP-SHAP iteratively removes nodes according to random permutations while preserving tree connectivity. After each perturbation, the resulting change in the model output is measured, and the SHAP value of each node is estimated as its average marginal contribution across permutations using permutation-based SHAP attribution [8]. The root node v_0 is always retained to ensure structural validity. In practice, we evaluate both forward and reverse permutations to improve estimation quality, while batching and caching model evaluations to maintain computational efficiency.

TABLE I: Open datasets used for training and evaluation.

Dataset	# Websites	# Fraudulent	# Benign	Year
phishpedia (train)	59 447	29 496	29 951	2021
knowPhish (test)	10 000	5000	5000	2023
δphish	6523	1012	5511	2017
zenodo (subset)	4000	2000	2000	2021

TABLE II: Performance on the *knowPhish* dataset.

Detector	ACC	F1	P	R	Time
HTML-GNN [14]	84.7	84.7	84.8	84.7	40 ms
GNN$^+$ [15]	88.4	88.4	88.7	88.4	30 ms
SpecularNet [1]	**93.9**	**93.9**	**94.0**	**93.9**	**20 ms**

III. Experimental Evaluation & Discussion

We evaluate DOM-SP-SHAP on SpecularNet using both real-world and adversarial phishing datasets summarized in Table I. SpecularNet is trained on the *phishpedia* dataset [9], containing approximately 59K webpages collected from Alexa [10] and OpenPhish [11]. Evaluation is conducted on the more recent *knowPhish* dataset [12], as well as the adversarial δPhish and zenodo SpacePhish datasets [13], which contain structurally manipulated phishing webpages generated through realistic HTML-level perturbations.

This setup enables both temporal and out-of-distribution evaluation: the model is trained only on the oldest dataset (*phishpedia*, 2021), while all testing is performed on newer and adversarial datasets spanning different phishing campaigns and structural manipulation strategies.

Table II summarizes SpecularNet's phishing-detection performance on the *knowPhish* dataset compared to recent graph-based phishing detectors. Specifically, we compare against HTML-GNN [14] and GNN$^+$ [15] using the implementations and model configurations provided by the respective authors, under the same training and evaluation setup as SpecularNet. The results show that SpecularNet achieves state-of-the-art detection performance while maintaining very low inference latency, outperforming previous GNN-based phishing detectors and demonstrating strong robustness under temporal distribution shift.

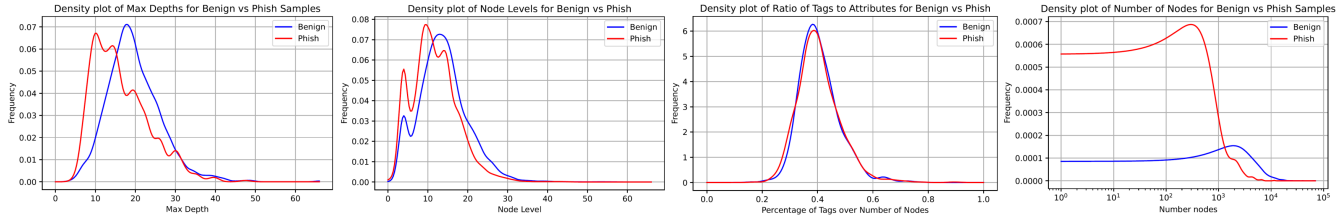
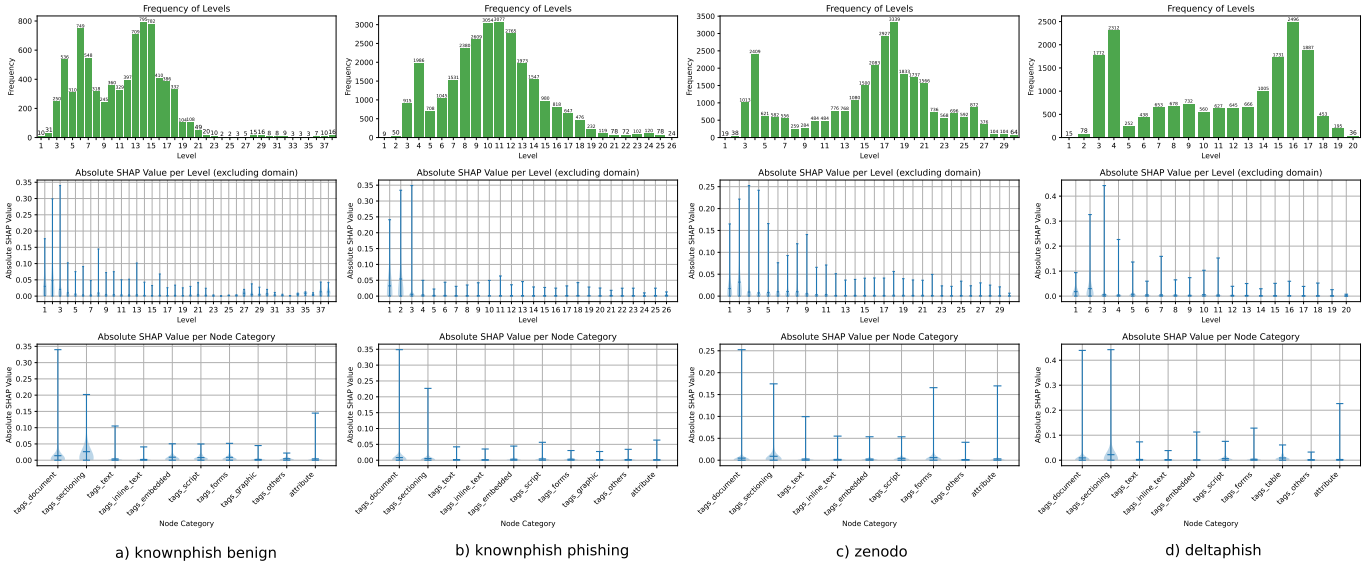


Fig. 1: Structural properties of *knowPhish* webpages. Benign webpages exhibit higher structural complexity, with greater maximum depth, deeper node-level distributions, and a larger number of DOM nodes, reflecting richer and more hierarchical layouts.



a) knownphish benign b) knownphish phishing c) zenodo d) deltaphish

Fig. 2: SpecularNet DOM-SP-SHAP results across datasets. Predictions are driven by structural layout invariants, with highest importance at shallow DOM levels. This explains strong generalization across phishing campaigns, while highlighting potential limitations under high-fidelity structural mimicry.

To understand SpecularNet's decision process, we apply DOM-SP-SHAP to generate node-level explanations over DOM trees. Fig. 1 shows that *knowPhish* benign webpages generally exhibit richer and deeper DOM structures, with larger node counts and more complex layouts, whereas phishing pages tend to be structurally simpler and concentrated near the HTML root.

Fig. 2 presents DOM-SP-SHAP results across the *knowPhish*, zenodo, and δPhish datasets. A highly consistent attribution pattern emerges across all datasets: predictive importance concentrates at shallow DOM levels, particularly depths 2–5, corresponding to the global layout structure of webpages, including major containers, navigation regions, and interaction areas. In contrast, deeper nodes contribute minimally despite being significantly more numerous.

Analysis by node category further reinforces this behavior. The highest attribution values are consistently associated with structural and sectioning tags defining page hierarchy and layout, while inline text, embedded media, and script-related nodes exhibit comparatively low importance. Form-related nodes contribute moderately, particularly in phishing pages, reflecting the role of credential-capture interfaces. Importantly, these attribution patterns remain stable even in adversarial datasets containing structurally manipulated webpages.

DOM-SP-SHAP evaluation indicates that SpecularNet primarily relies on structural layout invariants rather than superficial textual, visual, or script-based cues. This behavior aligns naturally with the architecture of the hierarchical graph autoencoder, where nodes near the DOM root act as structural bottlenecks during message passing and reconstruction. Overall, DOM-SP-SHAP helps explain the strong robustness and cross-dataset generalization achieved by SpecularNet, as webpage layout patterns tend to persist across phishing campaigns and are harder to manipulate without degrading visual fidelity.

Regarding limitations, DOM-SP-SHAP presents several challenges and opportunities for future work. The permutation-based attribution process remains computationally expensive for large DOM trees, and the current framework focuses exclusively on structural explanations, without considering visual rendering or dynamic webpage behavior. Future work includes improving attribution efficiency, extending the framework to multimodal webpage representations, and evaluating robustness against highly sophisticated phishing pages employing near-perfect structural mimicry.

ACKNOWLEDGMENTS

This work has been supported by the FWF Austrian Science Fund, Project reference I-6653, as well as by the Agencia Estatal de Investigacion, Project reference PCI2023-145974-2, as part of the EU CHIST-ERA-2022-SPiDDS-02 project GRAPHS4SEC. This work is also supported by the Catalan Institution for Research and Advanced Studies (ICREA Academia).

REFERENCES

[1] T. Song, P. Casas, and M. Meo, "Phishing the Phishers with SpecularNet: Hierarchical Graph Autoencoding for Reference-Free Web Phishing Detection," *arXiv preprint, arXiv:2603.01874*, 2026.

[2] ——, "Do We Really Need Reference-Based Phishing Detection? Unleashing the Power of GNN," in *2025 9th Network Traffic Measurement and Analysis Conference (TMA)*, 2025, pp. 1–4.

[3] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You?: Explaining the Predictions of Any Classifier," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2016, pp. 1135–1144.

[4] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[5] Z. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec, "GNNExplainer: Generating Explanations for Graph Neural Networks," in *Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.

[6] M. Vu, M. T. Thai, and V. Nguyen, "Probabilistic Graphical Model Explanations for Graph Neural Networks," in *Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 12 225–12 236.

[7] H. Yuan, H. Yu, J. Wang, K. Li, and S. Ji, "On Explainability of Graph Neural Networks via Subgraph Explorations," in *Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 12 241–12 252.

[8] R. Mitchell, J. Cooper, E. Frank, and G. Holmes, "Sampling Permutations for Shapley Value Estimation," *J. Mach. Learn. Res.*, vol. 23, no. 43, pp. 1–46, 2022.

[9] Y. Lin, R. Liu, D. M. Divakaran, J. Y. Ng, Q. Z. Chan, Y. Lu, Y. Si, F. Zhang, and J. S. Dong, "Phishpedia: A Hybrid Deep Learning Based Approach to Visually Identify Phishing Webpages," in *USENIX Secur. Symp.*, 2021, pp. 3793–3810.

[10] I. Alexa Internet. (2022) Alexa ranking. Accessed: October 2020. [Online]. Available: https://www.alexa.com/siteinfo

[11] OpenPhish. (2022) Openphish - phishing intelligence feed. Accessed: 3 June 2025. [Online]. Available: https://openphish.com

[12] Y. Li, C. Huang, S. Deng, M. L. Lock, T. Cao, N. Oo, H. W. Lim, and B. Hooi, "KnowPhish: Large Language Models Meet Multimodal Knowledge Graphs for Enhancing Reference-Based Phishing Detection," in *USENIX Secur. Symp.*, 2024, pp. 793–810.

[13] G. Apruzzese, M. Conti, and Y. Yuan, "SpacePhish: The Evasion-Space of Adversarial Attacks Against Phishing Website Detectors Using Machine Learning," in *Annu. Comput. Secur. Appl. Conf. (ACSAC)*, 2022, pp. 171–185.

[14] L. Ouyang and Y. Zhang, "Phishing Web Page Detection with HTML-Level Graph Neural Network," in *IEEE Int. Conf. Trust, Secur. Privacy Comput. Commun. (TrustCom)*, 2021, pp. 952–958.

[15] Y. Luo, L. Shi, and X.-M. Wu, "Unlocking the Potential of Classic GNNs for Graph-Level Tasks: Simple Architectures Meet Excellence," *arXiv preprint, arXiv:2502.09263*, 2025.