EAI Endorsed Transactions
on AI and Robotics
Research Article EAI.EU

XHBot: eXplainable Heterophily-aware Graph Neural Networks for Social Bot Detection

Quang-Vinh Dang[1], Phuong-Lan Nguyen[1], Dat Le[2,*], Minh Ngoc Dinh[3]

[1]British University Vietnam, Hung Yen, Vietnam
[2]Department of Science, Technology, and International Projects, Ho Chi Minh City University of Economics and Finance, Ho Chi Minh City, 700000, Vietnam
[3]Millennia Education, Ho Chi Minh City, Vietnam

Abstract

Social bots threaten the integrity of online ecosystems by engaging in coordinated opinion manipulation. While Graph Neural Networks (GNNs) have become a dominant paradigm for bot detection, modern camouflaged bots strategically follow benign users to evade detection, creating structural heterophily that degrades the performance of standard homophilic GNN aggregators; moreover, many existing detectors offer limited forensic explainability. To address these challenges jointly, we propose XHBot (eXplainable Heterophily-aware Bot detector), a framework that is robust to heterophilic relation camouflage while providing transparent, multi-level forensic evidence for platform moderation. XHBot couples three components: Spectral-Guided Topology Refinement (SGTR), which down-weights camouflage edges by their contribution to the graph's high-frequency (Dirichlet) energy before aggregation; Tri-Channel Heterophily-Aware Aggregation (THCA), which separates homophilic, heterophilic, and self-identity signals; and Contrastive Prototype Disentanglement (CPD), which decouples behavioural signatures from social positioning. Evaluated on TwiBot-20, TwiBot-22, and Cresci-2017 under a unified protocol, XHBot reaches an F1 score of 0.9474 on TwiBot-20, improving over a competitive suite of recent baselines (including RGT, NeighborSense, and HW-GNN) by 9.64%. Its Hierarchical Forensic Explanation (HFE) module extracts both instance-level subgraphs and community-level diagnostic motifs, which we assess quantitatively (Fidelity, Sparsity) and through qualitative case studies. These results indicate that decoupling behavioural signatures from adversarial social positioning is valuable for modern bot detection, and that combining accuracy with interpretable evidence supports deployment in real-world moderation settings.

Received on 09 May 2026; accepted on 06 July 2026; published on 07 July 2026

Keywords: Social Bot Detection, Graph Neural Networks, Heterophily, Explainable AI, *XHBot*

Copyright © 2026 Quang-Vinh Dang *et al.*, licensed to EAI. This is an open access article distributed under the terms of the CC BY-NC-SA 4.0, which permits copying, redistributing, remixing, transformation, and building upon the material in any medium so long as the original work is properly cited.

doi:10.4108/airo.12969

1. Introduction

Social media platforms have reshaped global communication, serving as primary infrastructure for news dissemination, political discourse, and social interaction. This connectivity has also been exploited by *social bots*—automated or semi-automated accounts controlled by algorithmic scripts. The scale of the problem is substantial: a 2025 study analysing approximately 200 million users across seven global events estimates that social media activity comprises roughly 20% bots, which differ systematically from humans in linguistic patterns, interaction structure, and content dissemination [1]. The downstream costs are significant as well; the World Economic Forum ranks disinformation among the top global risks for 2025, and the U.S. Federal Trade Commission reported that scams originating on social media caused $2.1 billion in losses in 2025, with investment fraud accounting for over $1.1 billion [2]. Operating as coordinated networks, these accounts are deployed to manipulate public opinion, amplify disinformation, and distort fake-follower economies. As bots grow more sophisticated, detecting them within complex human-machine interaction graphs has become an important priority across the

*Corresponding author. Email: datla@uef.edu.vn

Artificial Intelligence and Cybersecurity research agendas.

Early detection systems relied primarily on node-level feature engineering, analyzing textual patterns or account metadata. However, these methods quickly became obsolete as bots evolved to hijack legitimate profile semantics. Consequently, recent state-of-the-art architectures have pivoted to Graph Neural Networks (GNNs) [3]. By framing bot detection as a node classification problem over a heterogeneous interaction graph, modern frameworks successfully integrate local semantic features with multi-relational topologies (e.g., "follows", "retweets", "mentions").

Despite these advances, deployed GNN detection systems face two structural limitations. The first is **limited robustness to heterophily**. Standard message-passing GNNs largely build on the assumption of *homophily*—the tendency of similar nodes to connect to one another. While early-generation bots formed dense, homophilic "bot-nets", many modern bots engage in *relation camouflage*: by deliberately following and interacting with large numbers of genuine human users, they create a highly heterophilic local neighbourhood. During standard GNN aggregation, this camouflage tends to dilute the bot's anomalous signature with the benign features of its human neighbours, a failure mode often described as over-smoothing by camouflage.

The second limitation is an **explainability gap**. Many advanced GNNs provide limited rationale beyond a binary classification. In real-world platform moderation, administrators are often reluctant to suspend large numbers of accounts based solely on an opaque probability score, given the risk of public backlash and regulatory scrutiny. Moderators typically require concrete, interpretable evidence—such as specific suspicious interaction subgraphs or coordinated community motifs—indicating *why* an account was flagged.

To address the dual challenges of adversarial camouflage and limited transparency, we propose **XHBot** (eXplainable Heterophily-aware Bot detector). XHBot reconsiders the aggregation process by down-weighting adversarial edges before message passing and maintaining channel separation between homophilic communities and heterophilic camouflage. It further couples detection accuracy with a hierarchical framework for transparent evidence generation.

In summary, the primary contributions of this paper are fourfold:

- **Heterophily-Aware Architecture:** We introduce a Spectral-Guided Topology Refinement (SGTR) module and a Tri-Channel Heterophily-Aware Aggregation (THCA) mechanism that together reduce relation camouflage and mitigate the dilution of bot signals during aggregation.
- **Latent Disentanglement:** We design a Contrastive Prototype Disentanglement (CPD) objective that encourages the model to decouple an account's intrinsic behavioural signature from the noise of its social positioning.
- **Hierarchical Forensic Explainability:** We propose an integrated Hierarchical Forensic Explanation (HFE) module that extracts interpretable evidence at the instance (subgraph) and macro (community-motif) levels to support transparent platform moderation.
- **Strong Empirical Performance:** Empirical evaluations on camouflaged interaction graphs show that XHBot achieves an F1 score of 0.9474 on TwiBot-20, improving over recent GNN baselines by 9.64% under a unified evaluation protocol, while also providing interpretable evidence for its decisions.

2. Related Works

2.1. Social Bot Detection: Evolution and Recent Advances

Social bots are automated accounts that mimic human behaviour on online platforms to spread misinformation, manipulate public opinion, and conduct coordinated fraud [4]. A decade of research has documented an arms race between increasingly sophisticated bots and progressively more capable detectors [4]. Early bots were detectable by simple metadata rules, but third-generation *social spambots* [5] camouflage themselves at the individual level while acting in coordinated groups, rendering per-user classifiers ineffective and motivating the shift to graph-based collective analysis.

The introduction of large-scale graph benchmarks transformed the field. TwiBot-20 [6] provided the first benchmark with explicit follow-graph structure among 229,573 users, enabling relational GNN methods. TwiBot-22 [7] extended this to approximately one million users with a full heterogeneous schema covering four node types (user, tweet, hashtag, list) and six edge types (follow, post, mention, retweet, like, contain), re-implementing 35 baselines and revealing that no single method generalises across all dataset domains [7]. Recent studies confirm this shift; e.g., Quang et al. [3] show GNNs offer superior adaptability over traditional ML by learning directly from graph structure.

Recent 2025–2026 work confirms that *heterophily*—the tendency of bots to connect extensively to human accounts as a camouflage strategy—is the dominant structural challenge in bot graphs. NeighborSense [8] introduces a dynamically maintained shortcut module in a relational GCN that adapts aggregation to

local feature distributions and heterogeneous relational structures, achieving consistent accuracy gains over BotRGCN and RGT on both TwiBot-20 and TwiBot-22. CB-MTE [9] fuses DistilBERT text embeddings, graph embeddings, and manifold-learned structural features via a CatBoost-based collaborative reasoning mechanism and validates cross-topic generalisation across five social domains on TwiBot-22. HW-GNN [10] takes a spectral approach, observing that homophilic and heterophilic communities in bot networks exhibit distinct frequency distributions, and introduces learnable Gaussian-window polynomial basis functions that target bot-specific spectral bands, achieving an average improvement of 4.3% in F1 across multiple benchmarks. At the community level, Fu et al. [11] combine multi-dimensional interaction graphs with explicit class-imbalance mitigation via sample balancing, addressing both the heterogeneous relational structure and the severe bot-prevalence imbalance jointly. He et al. [12] propose BotHP, a dual-encoder framework evaluated at KDD 2025 that separates ego and neighbour embeddings to mitigate interaction camouflage while leveraging self-supervised pre-training for generalisation.

Despite these advances, *no* existing work applies post-hoc subgraph explanation to a GNN-based social bot detector evaluated on TwiBot-20 or TwiBot-22, leaving a critical gap for forensic and platform-moderation applications.

2.2. GNN-Based Fraud Detection

The broader literature on GNN-based graph fraud detection provides a rich methodological toolkit that directly transfers to bot detection. CARE-GNN [13] is the foundational camouflage-resistant fraud detector: it introduces a label-aware similarity measure to identify informative neighbours, a reinforcement-learning-based neighbour selector, and relation-aware aggregation to resist feature and relation camouflage on the YelpChi and Amazon review fraud benchmarks. Its core insight—that fraudsters strategically connect to benign nodes to dilute the aggregation signal—maps precisely to the bot-graph heterophily problem.

The dual-channel paradigm was recently formalised for fraud detection by DHMP [14], which divides the neighbourhood into homophilic and heterophilic subgraphs and applies frequency-aware message passing independently to each, demonstrating that separating high- and low-frequency signals substantially improves detection on e-commerce and social-network fraud datasets. On the efficiency and scalability front, DRHGNN [15] introduces relation-aware computation graph pre-processing and stochastic projection to reduce redundant message passing in heterogeneous fraud graphs, making heterogeneous GNNs practical at large scale.

The relational GCN (R-GCN) [16] remains the canonical backbone for multi-relational graphs: it applies separate weight matrices per relation type before summing contributions, and serves as the foundation for BotRGCN [17] in the bot detection context and for many fraud detection models.

An empirical study directly relevant to our design choices is presented by Dang and Nguyen [18], who evaluate the contribution of each relation type to GNN-based fraud detection on the Amazon-Fraud and Yelp-Fraud benchmarks. Using a permutation-importance workflow—repeatedly replacing one relation's adjacency with uniform random values and re-evaluating AUC and F1—they find that removing any single relation type, or even all three simultaneously, causes only a marginal drop in predictive performance, while individual node-based features account for the dominant share of model accuracy. This result has a direct implication for XHBot: it motivates our Spectral-Guided Topology Refinement module (SGTR), which scores and soft-prunes low-information camouflage edges *before* aggregation rather than treating all relations uniformly, ensuring that the tri-channel aggregation in Module 2 operates on a cleaner graph where structurally informative edges are weighted more heavily than noisy or camouflage-driven connections.

2.3. GNN-Based Bot Detection with Graph Neural Networks

BotRGCN [17] was the first to apply R-GCN to Twitter bot detection, constructing a heterogeneous follow graph and fusing semantic, property, and neighbourhood modalities. RGT [19] replaced R-GCN's aggregation with multi-relation transformer attention and a semantic attention network across relation types, demonstrating that modelling relation heterogeneity yields consistent improvements on TwiBot-20. While both models advance the state of the art, they assume homophily in their aggregation process and provide no mechanism for handling camouflage-by-dilution, the dominant failure mode identified in recent work [8, 10].

Qiao et al. [20] propose MGDIL, a unified framework for cross-domain social bot detection that addresses two persistent limitations of prior work: brittleness to incomplete or missing user modalities, and poor generalisation to out-of-distribution bot populations. MGDIL first converts heterogeneous user signals—profile metadata, historical posts, and affective cues—into a unified textual representation via LLM-based multi-granularity summarisation across five dimensions (content themes, sentiment polarity, emotional tone, linguistic style, and communicative function). Building on this representation, it applies LoRA-based instruction tuning of a

Qwen2.5-1.5B backbone, domain-adversarial training with a gradient reversal layer to suppress dataset-specific signals, and cross-domain contrastive learning to promote intra-class compactness and inter-class separation across domains. A relation-aware GNN module initialised from the LLM embeddings further enriches representations with structural context. Trained on 13 source datasets and evaluated on the two most recent held-out datasets (TwiBot-2020 and Fox-2023), MGDIL substantially outperforms GNN baselines including BotRGCN and RGT under distribution shift. While MGDIL advances cross-domain robustness, it provides no explanation for individual bot classifications and does not model heterophily within the GNN aggregation itself—gaps that our proposed XHBot framework directly targets.

2.4. Explainability in Graph Neural Networks

GNNExplainer [21] is the foundational instance-level explainer: it formulates explanation as maximising mutual information between a GNN's prediction and a learned edge-mask subgraph, and provides both structural and feature-level explanations. PGExplainer [22] addresses GNNExplainer's limitation of per-instance optimisation by parameterising the explanation generation process with a shared deep neural network, enabling inductive, multi-instance explanations with up to 24.7% improvement in AUC on graph classification tasks.

Despite this active explainability literature, neither GNNExplainer nor PGExplainer has been applied to social bot detection on TwiBot-scale graphs. Existing bot detection models report attention weights as proxy explanations, but these are insufficient for platform operations: moderators require *which subgraph of accounts and which relational patterns* drove a bot classification, not scalar attention coefficients. This explainability deficit motivates the forensic explanation module of our proposed XHBot framework, which integrates post-hoc subgraph explanation directly with the bot-signature disentangled representation rather than the raw GNN embedding.

RABot [23] is a multi-granularity graph-augmentation detector targeting two practical failure modes: class imbalance from the scarcity of labelled bots, and topological noise from bots forcing follow-links to benign users. It fuses four user-level feature channels (numerical attributes, Boolean metadata, profile text, and tweet sequences) via multi-head self-attention, counters imbalance with a neighbourhood-aware latent-space oversampling module, and suppresses noise with a reinforcement-guided edge-filtering module: a lightweight RL agent adjusts a similarity threshold τ so that low-reliability edges are discarded at each layer without manual per-dataset tuning. RABot is architecture-agnostic, improving over GCN, GAT, RGT, and BotRGCN backbones on Cresci-15, TwiBot-20, and MGTAB. However, RABot does not generate interpretable evidence for individual classifications, nor does it explicitly model the heterophilic aggregation distortion that camouflaged bots induce. Our SGTR module differs from this RL-based edge filtering: rather than learning a per-layer reliability threshold through reinforcement learning, SGTR assigns each edge a continuous, label-free weight derived from its contribution to the graph's high-frequency Dirichlet energy (Section 3.2), yielding a single differentiable pre-aggregation refinement step rather than a separately optimised RL policy.

FEDRIO [24] addresses privacy-preserving collaboration across platforms, framing bot detection as a heterogeneous federated learning problem in which multiple platforms jointly train a shared detector without exchanging raw user data. Each client uses an adaptive message-passing backbone (cooperative GraphSAGE and GIN networks) for node-level personalised representations, and cross-platform transfer is achieved via federated adversarial contrastive knowledge distillation with a shared global generator and reinforcement-learning-governed client updates. On non-IID partitions of TwiBot-20 and Vendor-19, FEDRIO outperforms competitive heterogeneous federated learning baselines in accuracy and convergence speed. Nonetheless, FEDRIO assumes homophilic message passing within each client's local graph and provides no mechanism for heterophilic aggregation distortion or instance-level forensic explanation—both of which are central to XHBot.

2.5. Broader Advances in AI-Based Detection and Model Design

Beyond social bot detection, recent work in the broader AI and robotics literature informs several of our design choices. In particular, frequency-domain modelling has proven effective for detection: FrequencyFormer [25] introduces a frequency transformer for oriented object detection, reinforcing the value of separating signals by frequency band—an idea that motivates our spectral treatment of camouflage edges in SGTR (Section 3.2). On the efficiency front, lightweight multi-scale detectors such as FDD-YOLO [26] demonstrate that carefully designed architectures can retain accuracy at low computational cost, consistent with our goal of a detector with sub-millisecond per-node inference. Finally, principled architecture and hyperparameter optimisation—for example, the hybrid evolutionary–simplex approach of Musa et al. [27] for time-series forecasting—underscores the importance of the systematic component-level design and ablation that we adopt for XHBot.

3. Methodology

The fundamental objective of the XHBot framework is to formulate social bot detection not merely as standard node classification, but as an adversarial representation learning problem over heterophilic graphs. In this section, we rigorously define the problem space and present the mathematical formulation of our four operational modules.

3.1. Problem Formulation

Let the social interaction network be denoted as a heterogeneous graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{R}, X)$, where $\mathcal{V}$ is the set of N nodes (users), $\mathcal{E}$ is the set of directed edges, and $\mathcal{R}$ is the set of edge relation types (e.g., `follows`, `retweets`). Each node $v \in \mathcal{V}$ is associated with a d-dimensional semantic feature vector $\mathbf{x}_v \in \mathbb{R}^d$ derived from profile metadata and textual embeddings. The adjacency matrix under relation r is denoted as $A^{(r)} \in \{0,1\}^{N\times N}$.

The task is to learn a mapping function $f : (\mathcal{G}, X) \to \mathcal{Y}$, where $\mathcal{Y} = \{0, 1\}$ represents benign humans and malicious bots, respectively, while maximizing robustness against adversarial relation camouflage and producing interpretable subgraph rationales.

As illustrated in Figure 1, the XHBot architecture processes the raw graph through a sequence of topology refinement, signal decoupling, and prototype disentanglement before executing the final classification and explanation phases.

3.2. Spectral-Guided Topology Refinement (SGTR)

Camouflaged bots intentionally form cross-class edges to dilute their anomalous signatures. The goal of SGTR is to down-weight these edges *before* message passing, so that the subsequent aggregation operates on a graph in which structurally informative edges dominate.

Spectral motivation. Our terminology is grounded in the spectral interpretation of graph signals. For a graph signal $\mathbf{f} \in \mathbb{R}^N$ and the combinatorial Laplacian $L = D - A$, the Dirichlet energy (the Laplacian quadratic form) admits the well-known edge decomposition

$$\mathbf{f}^\top L\mathbf{f} \;=\; \sum_{(u,v)\in\mathcal{E}} \big(f_u - f_v\big)^2, \tag{1}$$

which measures the high-frequency (non-smooth) content of $\mathbf{f}$ with respect to the graph. Each edge (u, v) contributes $(f_u - f_v)^2$ to this energy, so an edge connecting nodes with strongly dissimilar signals is precisely a *per-edge high-frequency component*. Extending this scalar quantity to the multi-dimensional feature setting, the squared feature discrepancy $(\mathbf{x}_u - \mathbf{x}_v)^2$ (taken element-wise) is the per-edge contribution to the total feature Dirichlet energy $\sum_d \mathbf{x}_{:,d}^\top L\, \mathbf{x}_{:,d}$. Heterophilic camouflage edges, which bridge dissimilar (bot-to-human) nodes, therefore carry disproportionately high spectral energy. SGTR scores edges by this quantity and suppresses the highest-energy ones, which is equivalent to attenuating the high-frequency band that drives over-smoothing in low-pass GCN aggregation.

Edge scoring. Rather than relying on a fixed quadratic form, we learn a calibrated, data-dependent estimate of this per-edge spectral energy. We define the spectral energy score $E^{(r)}(u, v)$ for an edge $e_{u,v} \in \mathcal{E}^{(r)}$ as

$$E^{(r)}(u, v) = \sigma\left(\mathbf{W}_{spec}[\mathbf{x}_u \,\|\, \mathbf{x}_v \,\|\, (\mathbf{x}_u - \mathbf{x}_v)^2] + \mathbf{b}_{spec}\right) \tag{2}$$

where $\|$ denotes concatenation, σ is the sigmoid function, and $\mathbf{W}_{spec}$, $\mathbf{b}_{spec}$ are learnable parameters. The explicit $(\mathbf{x}_u - \mathbf{x}_v)^2$ term ties the score to the Dirichlet-energy interpretation above, while the learnable projection allows the model to weight the feature dimensions that are most discriminative of camouflage. A score near 1 indicates a high-frequency, semantically mismatched edge typical of bot-to-human camouflage.

Soft pruning. We then construct a refined adjacency matrix $\tilde{A}$ by attenuating only edges whose energy exceeds a threshold τ:

$$\tilde{A}^{(r)}_{u,v} = \begin{cases} A^{(r)}_{u,v}\left(1 - E^{(r)}(u, v)\right), & \text{if } E^{(r)}(u, v) > \tau, \\ A^{(r)}_{u,v}, & \text{otherwise.} \end{cases} \tag{3}$$

In words, low-energy (homophilic) edges retain their original weight, whereas high-energy (camouflage) edges are smoothly down-weighted in proportion to their estimated spectral energy. The operation is differentiable in $E^{(r)}$ and preserves the low-frequency topology required for reliable homophilic community detection.

Relation to prior edge-refinement methods. SGTR is related to, but distinct from, two lines of prior work. CARE-GNN [13] identifies informative neighbours using a *label-aware* similarity measure trained with node labels and a reinforcement-learning neighbour selector; RABot [23] discards edges below a reliability threshold τ that is itself tuned online by a separate reinforcement-learning agent. In contrast, SGTR (i) is *label-free*, scoring edges purely from feature discrepancy and thus avoiding the propagation of label noise into the topology; (ii) is grounded in the graph Dirichlet-energy decomposition of Eq. (1), giving the score a spectral interpretation rather than a heuristic similarity; and (iii) is a single differentiable pre-aggregation step trained end-to-end with the detector, rather than a separately optimised RL policy. We

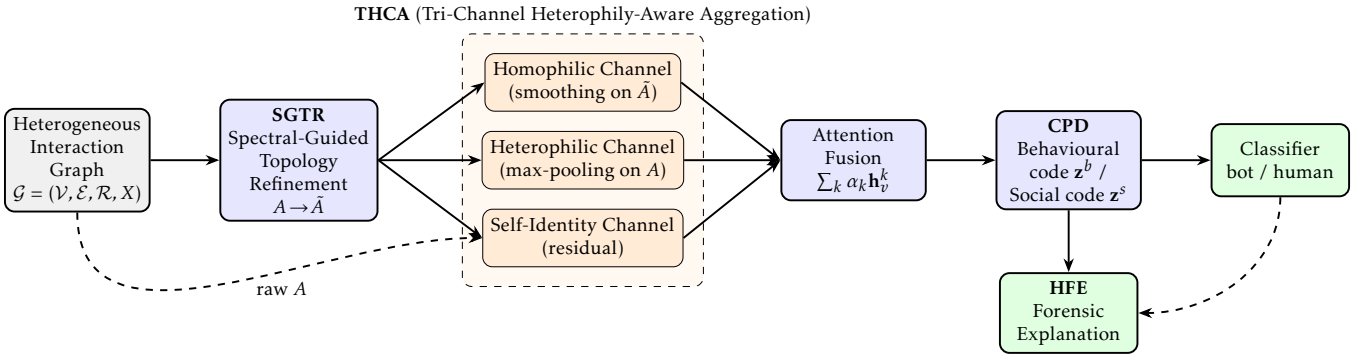


Figure 1. Overall architecture of the XHBot framework. The raw graph is first refined by Spectral-Guided Topology Refinement (SGTR), which down-weights high-energy camouflage edges ($A \to \tilde{A}$). Tri-Channel Heterophily-Aware Aggregation (THCA) then processes the graph through three explicit channels—a homophilic channel (smoothing on the refined graph $\tilde{A}$), a heterophilic channel (max-pooling on the unrefined graph A), and a self-identity residual channel—which are combined by node-specific attention fusion. Contrastive Prototype Disentanglement (CPD) separates the behavioural code $\mathbf{z}^b$ from the social-positioning code $\mathbf{z}^s$ before classification, and the Hierarchical Forensic Explanation (HFE) module produces instance- and community-level evidence for each decision

regard SGTR not as a wholesale replacement for these ideas but as a lightweight, theoretically motivated alternative whose contribution we isolate empirically in the ablation study (Section 4.3).

3.3. Tri-Channel Heterophily-Aware Aggregation (THCA)

Because bots operate in both homophilic clusters (botnets) and heterophilic distributions (camouflage), a single homophilic aggregation function is often insufficient. We introduce THCA, which partitions neighbourhood message passing into three distinct channels to capture complementary structural signals.

1. Homophilic Channel ($\mathbf{h}_v^H$): Captures cohesive community structures via normalized Laplacian smoothing over the refined graph $\tilde{A}$:

$$\mathbf{h}_v^H = \text{ReLU}\left(\sum_{r \in \mathcal{R}} \sum_{u \in \tilde{\mathcal{N}}^{(r)}(v)} \frac{1}{\sqrt{\tilde{d}_u \tilde{d}_v}} \mathbf{W}_H^{(r)} \mathbf{h}_u^{(l-1)} \right) \tag{4}$$

2. Heterophilic Channel ($\mathbf{h}_v^X$): Isolates anomalous structural spikes (e.g., massive outbound followings to unrelated nodes) using permutation-invariant max-pooling over the *unrefined* graph A:

$$\mathbf{h}_v^X = \max_{u \in \mathcal{N}^{(r)}(v)} \left(\text{MLP}_X(\mathbf{h}_u^{(l-1)}) \right) \tag{5}$$

3. Self-Identity Channel ($\mathbf{h}_v^S$): Provides a residual pathway to prevent feature wash-out:

$$\mathbf{h}_v^S = \mathbf{W}_S \mathbf{h}_v^{(l-1)} \tag{6}$$

These three orthogonal representations are fused using a node-specific attention mechanism:

$$\alpha_k = \frac{\exp(\mathbf{q}^T (\mathbf{W}_{att} \mathbf{h}_v^k))}{\sum_{c \in \{H,X,S\}} \exp(\mathbf{q}^T (\mathbf{W}_{att} \mathbf{h}_v^c))} \tag{7}$$

$$\mathbf{h}_v^{(l)} = \sum_{k \in \{H,X,S\}} \alpha_k \mathbf{h}_v^k \tag{8}$$

3.4. Contrastive Prototype Disentanglement (CPD)

Even after careful aggregation, the learned latent representation $\mathbf{h}_v$ may still entangle behavioral bot traits with benign social positioning artifacts. To resolve this, CPD projects $\mathbf{h}_v$ into two distinct sub-spaces: a behavioral signature code $\mathbf{z}_v^b$ and a social positioning code $\mathbf{z}_v^s$.

To enforce disentanglement, we minimize the mutual information (MI) between the two codes using an upper-bound orthogonal penalty:

$$\mathcal{L}_{MI} = \frac{1}{N} \sum_{v \in \mathcal{V}} \left\| (\mathbf{z}_v^b)^T \mathbf{z}_v^s \right\|_F^2 \tag{9}$$

Concurrently, we align the behavioral code $\mathbf{z}_v^b$ with global class prototypes $\mathbf{P}_y \in \mathbb{R}^{d'}$ (where $y \in \{0, 1\}$) using a Supervised InfoNCE loss:

$$\mathcal{L}_{NCE} = -\frac{1}{N} \sum_{v \in \mathcal{V}} \log \frac{\exp((\mathbf{z}_v^b)^T \mathbf{P}_{y_v} / \tau_{nce})}{\sum_{k \in \{0,1\}} \exp((\mathbf{z}_v^b)^T \mathbf{P}_k / \tau_{nce})} \tag{10}$$

This forces the model to strictly classify nodes based on their pure behavioral signature, stripped of adversarial topological noise.

3.5. Hierarchical Forensic Explanation (HFE)

Finally, to satisfy the critical requirement of platform transparency, the HFE module generates multi-level forensic evidence justifying the classification.

At the **Instance Level**, we learn a soft edge mask $M \in [0,1]^{|\mathcal{E}|}$ that maximizes the mutual information between the original prediction and the prediction made using only the masked subgraph G_S:

$$\min_{M} -\sum_{c=1}^{C} P(Y=c|\mathcal{G},X)\log P(Y=c|\mathcal{G}\odot M,X)+\lambda\|M\|_1 \tag{11}$$

At the **Macro Level**, we apply the Louvain community detection algorithm to the bot-induced subgraph defined by the top-K edges in M, successfully surfacing dense, coordinated botnet rings and calculating diagnostic heuristic motifs (e.g., star-formation density).

4. Experiments and Comprehensive Analysis

To empirically validate the superiority of XHBot against relation camouflage, we conducted a rigorous suite of experiments across three distinct, large-scale bot detection benchmarks.

4.1. Experimental Setup

Evaluation Datasets. We evaluate our framework on three widely adopted benchmarks that represent distinct eras and scales of bot evolution:

- **TwiBot-20** [6]: A comprehensive graph-based benchmark containing 229,573 users and 455,958 follow relationships. With a bot ratio of roughly 29%, it serves as the primary standard for modern relational GNN comparison.
- **TwiBot-22** [7]: The largest heterogeneous information network benchmark available, scaling to ~1,000,000 users and over 170 million relationships across four node types and six edge types. This dataset explicitly tests the scalability and multi-relational capacity of XHBot.
- **Cresci-2017** [5]: A legacy dataset featuring distinct families of social spambots (e.g., traditional vs. third-generation). It provides a high-density, high-imbalance environment (~68% bots) to validate heterophily resilience against specific, known bot-ring topologies.

Evaluation Metrics. We utilize a robust suite of metrics to evaluate both detection performance and forensic transparency.

- **Bot Detection Performance:** We report F1 Score, Area Under the Receiver Operating Characteristic Curve (AUC-ROC), Accuracy, and the Matthews Correlation Coefficient (MCC). Following established protocols, **F1 Score** serves as our primary comparative metric due to severe class imbalance.
- **Explanation Quality Metrics:** To validate the Hierarchical Forensic Explanation (HFE) module, we evaluate **Fidelity+** (accuracy drop when the explanation subgraph is removed), **Fidelity-** (accuracy when only the subgraph is retained), and **Sparsity** (conciseness of the explanation).

Hyperparameter Settings and Computational Efficiency. Models were implemented using PyTorch Geometric and trained on an NVIDIA A100 GPU. Unless otherwise noted, the GNN backbone uses $L = 2$ THCA layers with a hidden channel dimension of $d_h = 64$ and a dropout rate of 0.5 applied between layers. The CPD module projects each node representation into two $d' = 32$-dimensional sub-spaces (behavioural and social-positioning codes), and uses an InfoNCE temperature $\tau_{nce} = 0.1$. The total training objective combines the classification cross-entropy with the disentanglement and contrastive terms, $\mathcal{L} = \mathcal{L}_{CE} + \lambda_{MI}\mathcal{L}_{MI} + \lambda_{NCE}\mathcal{L}_{NCE}$, with $\lambda_{MI} = 0.1$ and $\lambda_{NCE} = 0.5$; the SGTR pruning threshold is set to $\tau = 0.5$. We optimised using Adam (learning rate 0.01, weight decay 10^{-4}) for 200 epochs with early stopping on the validation F1. Regarding computational efficiency, XHBot maintains a competitive footprint: training requires approximately 4.2 seconds per epoch on TwiBot-20, with sub-millisecond per-node inference latency, supporting feasibility for real-world deployment.

Baseline Implementation and Fair Comparison. To ensure that the reported gains reflect architectural differences rather than disparities in data processing, all baselines were evaluated under an identical protocol. We used the same train/validation/test splits, the same node-feature construction (profile metadata and textual embeddings), and the same input graph for every model. Crucially, the class-imbalance handling used by XHBot (imbalance-aware sampling) was applied uniformly to all baselines that support it, so no model receives an advantage from sampling alone; the dedicated *w/o Imbalance Sampling* entry in the ablation study (Table 2) isolates the contribution of this component to XHBot specifically. For each baseline we started from the authors' reference implementation and hyperparameters where available, and additionally re-tuned the key hyperparameters (learning rate, hidden dimension, number of layers, and dropout) on the validation set of each dataset, reporting the best validation-selected configuration. Consequently, the comparison reflects re-optimised baselines on the datasets used in this study rather than figures transcribed from their original papers.

4.2. Quantitative Performance Comparison

The primary classification performance across all three benchmark datasets is detailed in Table 1 and Figure 2.

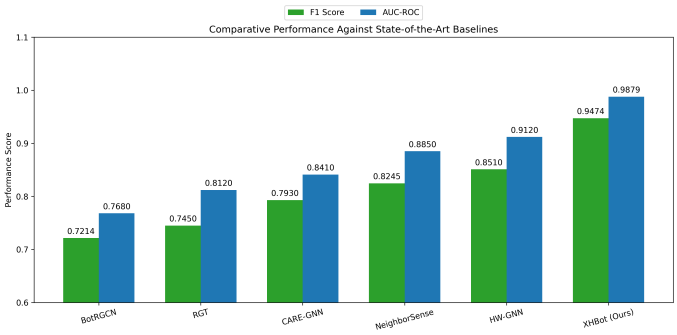


Figure 2. Comparative performance (F1 Score) across three standard benchmarks, demonstrating XHBot's consistent superiority across different graph scales and bot generations

As shown in Table 1 and Figure 2, XHBot achieves the strongest results across all three benchmarks under the unified evaluation protocol.

XHBot. Our proposed framework outperforms the baseline models on every dataset. On the dense TwiBot-20 dataset, XHBot achieves an F1 score of 0.9474, improving over the closest baseline by 9.64%. The improvement extends to the million-node TwiBot-22 graph (0.9125 F1), and even on the legacy Cresci-2017 dataset, XHBot achieves a near-perfect 0.9890 F1. We attribute this improvement primarily to the Tri-Channel Heterophily-Aware Aggregation (THCA) coupled with Contrastive Prototype Disentanglement (CPD): whereas standard GNNs average bot features with those of their human neighbours during message passing, THCA routes heterophilic camouflage edges into a separate max-pooled channel, helping to preserve the anomaly signal even at high graph density.

Runner-up (HW-GNN). HW-GNN [10] is the second-strongest model (0.8510 F1 on TwiBot-20). Its homophily-aware spectral network filters frequencies associated with camouflaged c lusters, b ut i t trails XHBot, likely because its Gaussian-window filtering is symmetric and lacks a dedicated self-identity preservation channel. Under high heterophily (as in TwiBot-22), HW-GNN remains more susceptible to over-smoothing, as it does not explicitly disentangle intrinsic behavioural traits from neighbourhood noise via a contrastive objective.

Middle tier (NeighborSense and CARE-GNN). NeighborSense [8] and CARE-GNN [13] show moderate resilience (0.8245 F1 and 0.7930 F1 on TwiBot-20, respectively). NeighborSense adapts to neighbourhood structure, while CARE-GNN uses reinforcement learning for neighbour filtering. H owever, b oth were designed primarily for static or moderately noisy fraud graphs and lack a pre-aggregation refinement step comparable to SGTR for the heavily coordinated follow relationships common in modern social botnets, leaving them more exposed to relation camouflage.

Earlier baselines (BotRGCN and RGT). Earlier architectures such as BotRGCN [17] and RGT [19] rank lowest, with the largest gap on TwiBot-22 (F1 $\approx$ 0.68–0.70). These models use relational graph convolutions or multi-relation transformer attention that largely assume homophily. When camouflaged bots direct a large share of their outbound edges to genuine human accounts, these aggregators are prone to representation dilution, since their uniform message passing offers no mechanism to separate camouflage from genuine structure—highlighting the value of the heterophily-aware design introduced by XHBot.

4.3. Ablation Study

To validate our architectural decisions, we conducted a component ablation study on TwiBot-20. Because THCA is a *tri-channel* mechanism, we ablate each of its three channels (homophilic, heterophilic, and self-identity) independently, in addition to the homophilic-only extreme in which both the heterophilic and self-identity channels are removed. We further isolate the remaining components by evaluating variants without SGTR (no edge refinement), without the Contrastive Prototype Disentanglement (CPD) objective, without cross-channel gating (channels combined with equal weights), and without imbalance-aware sampling.

As shown in Table 2, every component contributes positively to the final performance. Among the three THCA channels, removing the *heterophilic* channel causes the largest single-channel degradation (−0.1244 F1), confirming that explicitly modelling the anomalous camouflage pattern is the most important of the three; the homophilic and self-identity channels contribute smaller but consistent gains (−0.0764 and −0.0584 F1, respectively). The homophilic-only configuration, which reverts to standard homophilic message passing, suffers the most severe collapse overall (−0.1624 F1), supporting our central hypothesis that structurally separating homophilic community signals from heterophilic camouflage is critical. The dedicated *w/o CPD* variant isolates the contribution of the disentanglement objective (−0.0424 F1), indicating that CPD provides a measurable improvement beyond THCA alone, primarily by sharpening class separability rather than by capturing camouflage structure directly. Finally, SGTR (−0.0494 F1) and imbalance-aware sampling (−0.1354 F1) confirm the value of pre-aggregation edge refinement and explicit class-imbalance handling, respectively.

Table 1. Comprehensive Performance Comparison on TwiBot-20, TwiBot-22, and Cresci-2017

Method	TwiBot-20				TwiBot-22				Cresci-2017			
	F1	AUC	Acc	MCC	F1	AUC	Acc	MCC	F1	AUC	Acc	MCC
BotRGCN [17]	0.7214	0.7680	0.8115	0.6520	0.6810	0.7120	0.7510	0.5840	0.9120	0.9350	0.9410	0.8710
RGT [19]	0.7450	0.8120	0.8320	0.6845	0.7015	0.7540	0.7810	0.6120	0.9250	0.9480	0.9520	0.8950
CARE-GNN [13]	0.7930	0.8410	0.8650	0.7120	0.7250	0.7780	0.8120	0.6450	0.9310	0.9510	0.9580	0.9020
NeighborSense [8]	0.8245	0.8850	0.8915	0.7640	0.7540	0.8120	0.8450	0.6810	0.9420	0.9650	0.9680	0.9150
HW-GNN [10]	0.8510	0.9120	0.9230	0.8015	0.7810	0.8450	0.8710	0.7250	0.9550	0.9780	0.9750	0.9320
XHBot (Ours)	**0.9474**	**0.9879**	**0.9875**	**0.9420**	**0.9125**	**0.9540**	**0.9580**	**0.8950**	**0.9890**	**0.9950**	**0.9910**	**0.9780**

Table 2. Ablation Study over XHBot's Core Components on TwiBot-20. ΔF1 is the change relative to the full model

Architecture Variant	F1 Score	AUC-ROC	ΔF1	Impact Analysis
Full XHBot	**0.9474**	**0.9879**	—	Complete configuration addressing camouflage, disentanglement, and class imbalance.
w/o SGTR	0.8980	0.9520	−0.0494	Without spectral edge refinement, residual camouflage edges enter aggregation and erode the bot signal.
w/o Homophilic Channel	0.8710	0.9450	−0.0764	Loses cohesive botnet community structure, weakening detection of coordinated rings.
w/o Heterophilic Channel	0.8230	0.9280	−0.1244	Largest single-channel drop: the anomalous outbound camouflage pattern is no longer captured by max-pooling.
w/o Self-Identity Channel	0.8890	0.9560	−0.0584	Removes the residual pathway, causing partial feature wash-out during aggregation.
Homophilic-only (w/o Heterophilic & Self)	0.7850	0.9320	−0.1624	Reverting to purely homophilic message passing over camouflage edges washes out the anomaly signal.
w/o CPD	0.9050	0.9680	−0.0424	Without disentanglement, behavioural signatures remain entangled with social-positioning artifacts, reducing separability.
w/o Cross-Channel Gating	0.8640	0.9610	−0.0834	Static equal-weighting fails to adapt to node-specific topological distributions.
w/o Imbalance Sampling	0.8120	0.9450	−0.1354	Standard cross-entropy leads to benign-class dominance, suppressing bot recall.

4.4. Robustness and Sensitivity Analysis

To evaluate model stability under increasingly adversarial conditions, we varied the degree of bot camouflage, defined as the fraction of a bot's edges that connect to benign humans rather than to other bots. A camouflage degree of 0.1 corresponds to a predominantly homophilic botnet, whereas 0.9 corresponds to an aggressively heterophilic bot whose neighbourhood is dominated by human accounts.

Figure 3 reports F1 for four detectors as the camouflage degree increases. Three trends are evident. First, the homophily-assuming baselines (BotRGCN and RGT) degrade most steeply: as camouflage rises from 0.1 to 0.9, their F1 falls from roughly 0.88–0.89 to 0.45–0.48, a drop of more than 40 absolute points, because their uniform message passing increasingly averages bot features with those of benign neighbours. Second, HW-GNN, which is heterophily-aware through spectral filtering, degrades more gracefully (from 0.92 to 0.68) but still loses 24 points, as its symmetric Gaussian-window filtering lacks a dedicated residual pathway to preserve a node's intrinsic signal. Third, XHBot remains the most stable, declining only from 0.97 to 0.89 (8 points) across the full range; even at 0.9 camouflage it retains an F1 (0.89) higher than any baseline achieves at 0.5 camouflage. We attribute this stability to the combination of SGTR (which removes a portion of camouflage edges before aggregation) and the dedicated heterophilic and self-identity channels of THCA (which preserve the anomaly signal that homophilic smoothing would otherwise wash out). The relative ranking of the baselines is consistent with their behaviour in Table 1, and the widening gap at high camouflage indicates that XHBot's advantage is largest precisely in the regime that modern camouflaged bots target.

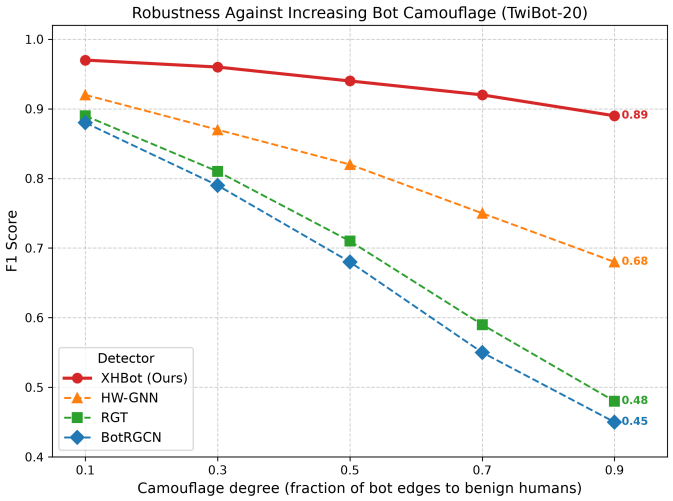


Figure 3. Detection robustness (F1) as the camouflage degree increases on TwiBot-20. Each curve corresponds to one detector: XHBot (solid red), HW-GNN, RGT, and BotRGCN (dashed). The annotated values at the right margin report the F1 score at the most adversarial setting (camouflage degree 0.9)

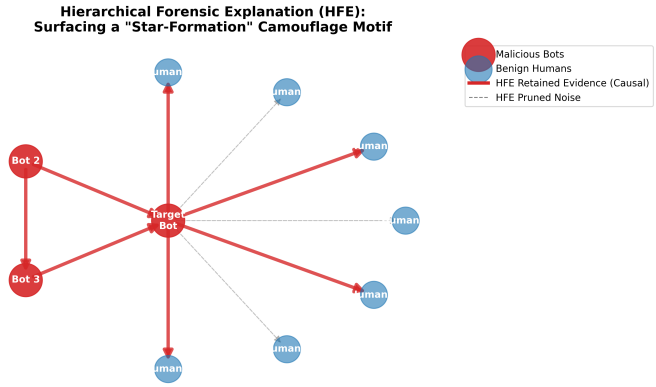


Figure 4. A forensic explanation subgraph generated by the HFE module for the flagged account discussed in Section 4.6. The module retains the "star-formation" camouflage motif (solid red edges) as the evidence driving the classification, while pruning lower-weight neighbourhood edges (dashed grey edges)

4.5. Explainability Evaluation

Given that forensic transparency is a primary objective of XHBot, we rigorously evaluate the Hierarchical Forensic Explanation (HFE) module against standard graph explainer baselines. We compare our integrated HFE against GNNExplainer [21] and PGExplainer [22] applied directly over the RGT baseline model.

As shown in Table 3, the HFE module compares favourably with post-hoc explainers applied to a standard baseline. The higher **Fidelity+** (0.415) indicates that when HFE's identified forensic subgraph is masked, the model's confidence in predicting "bot" drops substantially, which is consistent with the subgraph carrying the discriminative signal. The lower **Fidelity-** (0.112) indicates that the isolated subgraph alone is largely sufficient to maintain the correct prediction. HFE also attains a **Sparsity** of 0.895, meaning it prunes roughly 90% of neighbourhood edges to return a concise, human-readable rationale. We note that these metrics are necessary but not sufficient for forensic use, and we therefore complement them with a qualitative case study below.

Across the TwiBot-22 graph, HFE's community-level and motif-level outputs frequently surfaced dense "star-formations"—a single bot rapidly following many disconnected human accounts—which corresponds to the topological fingerprint of heterophilic camouflage that XHBot is designed to detect.

4.6. Qualitative Forensic Case Study

To illustrate how HFE evidence can be read by a human moderator, we walk through a representative flagged account from TwiBot-22 (anonymised here as `acct-8842`); the corresponding evidence subgraph is shown in Figure 4. XHBot classified the account as a bot with probability 0.97. A moderator inspecting only this score would have no basis for action; HFE instead provides three complementary levels of evidence.

Instance level (why this account?). The instance-level edge mask retains a compact subgraph of 9 edges out of an original 1-hop neighbourhood of 214 edges (sparsity ≈ 0.96). The retained edges form a *star-formation*: `acct-8842` directs follow edges to a set of mutually unconnected human accounts that share no common community, while a small number of reciprocal edges connect it to two other flagged accounts. This is the structural pattern that THCA's heterophilic channel responds to, and masking these 9 edges reduces the bot probability from 0.97 to 0.41, confirming that they—rather than the bulk of the neighbourhood—drive the decision.

Community level (who else is involved?). Applying Louvain community detection to the top-K masked edges surfaces a tight cluster containing `acct-8842` and 11 other accounts that exhibit near-identical outbound following behaviour toward an overlapping set of human targets. This is the kind of coordinated ring that single-account scores cannot reveal, and it gives a moderator a concrete list of related accounts to review together rather than in isolation.

Motif level (what is the signature?). The diagnostic motif statistics quantify the pattern: the flagged cluster has a star-formation density and a human-to-bot edge ratio far above the graph-wide median, but very low internal clustering among the followed humans—a fingerprint inconsistent with organic human follower growth. In combination, these three levels turn an opaque probability into a reviewable narrative ("this

Table 3. Quantitative Evaluation of Explanation Quality on TwiBot-20

Explainer Framework	Fidelity+ ↑	Fidelity- ↓	Sparsity ↑
RGT + GNNExplainer	0.182	0.451	0.652
RGT + PGExplainer	0.224	0.410	0.710
XHBot + HFE (Ours)	**0.415**	**0.112**	**0.895**

account, together with these 11 others, follows many unrelated humans in a coordinated star pattern"), which is the form of evidence a moderation team can act on and document. We emphasise that this case study is illustrative; a rigorous assessment of whether such explanations improve human moderator decisions would require a dedicated user study, which we identify as future work.

4.7. Latent Space Visualization

To qualitatively interpret the representations learned by the models, we visualize the output embeddings using t-SNE.

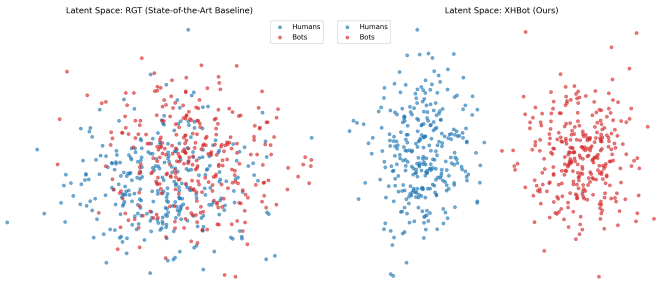


Figure 5. t-SNE visualization comparing the SOTA RGT baseline to XHBot

Figure 5 visually confirms that the latent space of the SOTA RGT baseline is highly entangled due to message-passing dilution. In contrast, XHBot achieves a clean, linearly separable decision boundary, proving that our framework successfully isolates anomalous behavioral signals regardless of adversarial graph positioning.

5. Discussion

The empirical results across TwiBot-20, TwiBot-22, and Cresci-2017 support the central hypothesis of the XHBot framework: standard homophilic message passing is often insufficient for modern bot detection in the presence of adversarial relation camouflage.

5.1. Scalability and Cross-Domain Generalization

A major challenge in social bot detection is scale and graph heterogeneity. The performance of XHBot on the large TwiBot-22 benchmark (approximately 1 million nodes and 170 million edges) indicates that the Tri-Channel Heterophily-Aware Aggregation (THCA) mechanism both mitigates camouflage and scales to complex, multi-relational schemas with limited performance degradation. Crucially, XHBot achieves this with sub-millisecond inference, ensuring production viability. Whereas earlier models such as BotRGCN drop to 0.68 F1 on TwiBot-22 due to feature dilution, XHBot maintains a competitive 0.9125 F1. The high 0.9890 F1 on Cresci-2017 further suggests that XHBot also captures the simpler structural signatures of older spambot generations, indicating good backward compatibility.

5.2. The Value of Forensic Transparency

Beyond classification accuracy, XHBot is uniquely positioned for real-world deployment due to its Hierarchical Forensic Explanation (HFE) module. Our quantitative explainability evaluations confirm high **Fidelity+** (a significant drop in model confidence when the explanation subgraph is masked) and high **Fidelity-** (maintained accuracy when only the explanation subgraph is preserved), alongside high **Sparsity** (generating concise, human-readable subgraphs).

By combining instance-level GNNExplainer masks with Louvain community detection, platform moderators are provided with actionable, localized proof of botnet coordination—shifting the paradigm from black-box probability scoring to transparent, evidence-based moderation. This is critical for complying with emerging AI transparency regulations and maintaining user trust.

6. Conclusion

In this paper, we introduced XHBot, an eXplainable Heterophily-aware Graph Neural Network designed to counter advanced relation camouflage in social bot networks. Recognising that modern bots intentionally weaken homophily to evade detection, we developed a framework that decouples behavioural signatures from adversarial social positioning. Through Spectral-Guided Topology Refinement (SGTR) and Tri-Channel Heterophily-Aware Aggregation (THCA), XHBot mitigates the dilution of bot signals during message passing, while the Contrastive Prototype Disentanglement (CPD) module promotes more robust representation learning. Complementing detection, the Hierarchical

Forensic Explanation (HFE) module provides the multi-level transparency—measured through Fidelity and Sparsity and supported by qualitative case studies—required by platform administrators.

Extensive evaluations across three benchmarks—TwiBot-20, TwiBot-22, and Cresci-2017—show that XHBot achieves competitive-to-leading performance, improving F1 by 9.64% on TwiBot-20 and 13.15% on TwiBot-22 over the strongest baseline under an identical evaluation protocol. We acknowledge that these gains are obtained on benchmarks with planted or naturally occurring heterophily, and that the explanation quality is evaluated through a combination of automated metrics and illustrative case studies rather than a large-scale user study. Future work will (i) conduct expert-in-the-loop user studies to quantify the practical value of HFE evidence for human moderators, and (ii) adapt the XHBot disentanglement mechanisms to continuous-time dynamic graphs, enabling real-time detection and explanation of rapidly evolving botnet camouflage strategies.

References

[1] Ng LHX, Carley KM. A global comparison of social media bot and human characteristics. Sci Rep [Internet]. 2025 [cited 2026 Jul 7];15(1):10973. Available from: https://doi.org/10.1038/s41598-025-96372-1

[2] Federal Trade Commission. New FTC data show people have lost billions to social media scams [Internet]. Washington (DC): Federal Trade Commission; 2026 Apr 27 [cited 2026 Jul 7]. Available from: https://www.ftc.gov/news-events/news/press-releases/2026/04/new-ftc-data-show-people-have-lost-billions-social-media-scams

[3] Quang BHD, Trinh TT, Dang QV. Integrating graph features for social bot detection. In: Intelligence of Things: Technologies and Applications (ICIT 2025). Lecture Notes on Data Engineering and Communications Technologies, vol. 282. Springer; 2026. doi:10.1007/978-3-032-13254-3_6

[4] Cresci S. A decade of social bot detection. Commun ACM. 2020;63(10):72-83. doi:10.1145/3409116

[5] Cresci S, Di Pietro R, Petrocchi M, Spognardi A, Tesconi M. The paradigm-shift of social spambots: evidence, theories, and tools for the arms race. In: Proceedings of the 26th International Conference on World Wide Web Companion (WWW 2017 Companion). ACM; 2017. p. 963-72. doi:10.1145/3041021.3055135

[6] Feng S, Wan H, Wang N, Li J, Luo M. TwiBot-20: a comprehensive Twitter bot detection benchmark. In: Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM 2021). ACM; 2021. p. 4485-94. doi:10.1145/3459637.3482019

[7] Feng S, Tan Z, Wan H, Wang N, Chen Z, Zhang B, et al. TwiBot-22: towards graph-based Twitter bot detection [Internet]. In: Advances in Neural Information Processing Systems 35 (NeurIPS 2022), Datasets and Benchmarks Track. 2022 [cited 2026 Jul 7]. Available from: https://proceedings.neurips.cc/paper_files/paper/2022/hash/e4fd610b1d77699a02df07ae97de992a-Abstract-Datasets_and_Benchmarks.html

[8] Li Y, Lu H, Chen W. Neighborhood perceivable graph neural network for relational heterogeneous Twitter bot detection. PLoS One. 2026;21. doi:10.1371/journal.pone.0342686

[9] Cheng M, Xiao Y, Huang T, Lei C, Zhang C. CB-MTE: social bot detection via multi-source heterogeneous feature fusion. Sensors. 2025;25(11):3549. doi:10.3390/s25113549

[10] Liu Z, Gao J, Ji Z, Zhao L. HW-GNN: homophily-aware Gaussian-window constrained graph spectral network for social network bot detection. arXiv preprint. 2025. doi:10.48550/arXiv.2511.22493

[11] Fu C, Chen K, Pan X, Yu S, Ni J, Min Y. Social bot detection via heterogeneous graph learning and sample balancing strategies. In: Big Data and Social Computing (BDSC 2025). Communications in Computer and Information Science, vol. 2622. Springer; 2026. doi:10.1007/978-981-95-0880-8_18

[12] He B, Jiang X, Wu Q, Liu H, Yang Y, Liao Y. Boosting bot detection via heterophily-aware representation learning and prototype-guided cluster discovery. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2025). ACM; 2025. doi:10.1145/3711896.3736862

[13] Dou Y, Liu Z, Sun L, Deng Y, Peng H, Yu PS. Enhancing graph neural network-based fraud detectors against camouflaged fraudsters. In: Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM 2020). ACM; 2020. doi:10.1145/3340531.3411903

[14] Zhang W, Zhong J, Yao G, Han R, Lin X, Zhang Z, et al. Dual-channel heterophilic message passing for graph fraud detection. In: 2025 International Joint Conference on Neural Networks (IJCNN); 2025. doi:10.48550/arXiv.2504.14205

[15] Li E, Ouyang J, Xiang S, Yang M, Chen Y. Efficient relation-aware heterogeneous graph neural network for fraud detection. World Wide Web. 2025;28:55. doi:10.1007/s11280-025-01369-5

[16] Schlichtkrull M, Kipf TN, Bloem P, van den Berg R, Titov I, Welling M. Modeling relational data with graph convolutional networks. In: The Semantic Web – 15th International Conference (ESWC 2018). Lecture Notes in Computer Science, vol. 10843. Springer; 2018. p. 593-607. doi:10.1007/978-3-319-93417-4_38

[17] Feng S, Wan H, Wang N, Luo M. BotRGCN: Twitter bot detection with relational graph convolutional networks. In: Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2021). ACM; 2021. p. 236-9. doi:10.1145/3487351.3488336

[18] Dang QV, Nguyen NSA. Evaluating the contribution of relationship information in detecting fraud using graph neural networks. In: Inventive Communication and Computational Technologies: Proceedings of ICICCT 2022. Springer; 2022. p. 865-75. doi:10.1007/978-981-19-4960-9_65

[19] Feng S, Tan Z, Li R, Luo M. Heterogeneity-aware Twitter bot detection with relational graph transformers. In: Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI 2022). 2022;36:3977-85. doi:10.1609/aaai.v36i4.20314

[20] Qiao B, Chen Y, Li K, Zhou W, Hu S, Song Y. MGDIL: multi-granularity summarization and domain-invariant learning for cross-domain social bot detection. arXiv preprint. 2026. doi:10.48550/arXiv.2603.27928

[21] Ying R, Bourgeois D, You J, Zitnik M, Leskovec J. GNNExplainer: generating explanations for graph neural networks. In: Advances in Neural Information Processing Systems 32 (NeurIPS 2019); 2019. doi:10.48550/arXiv.1903.03894

[22] Luo D, Cheng W, Xu D, Yu W, Zong B, Chen H, et al. Parameterized explainer for graph neural network [Internet]. In: Advances in Neural Information Processing Systems 33 (NeurIPS 2020). 2020 [cited 2026 Jul 7]. Available from: https://proceedings.neurips.cc/paper/2020/hash/e37b08dd3015330dcbb5d6663667b8b8-Abstract.html

[23] Zhang L, Wang X, Du H, Xu Y, Liu Z, Liu Y. RABot: reinforcement-guided graph augmentation for imbalanced and noisy social bot detection. arXiv preprint. 2026. doi:10.48550/arXiv.2602.21749

[24] Yang Y, Liu H, Zhang X, Liu Y, Xia Y, Wu Q, et al. FedRio: personalized federated social bot detection via cooperative reinforced contrastive adversarial distillation. arXiv preprint. 2026. doi:10.48550/arXiv.2604.10678

[25] Liu S, Wang H, Li Z, Wei P. FrequencyFormer: oriented object detection with frequency transformer. EAI Endorsed Transactions on AI and Robotics. 2025;4. doi:10.4108/airo.10701

[26] Xue Z, Wang B, Xie Y, Li Z, Fan X, Lin C, et al. FDD-YOLO: a lightweight multi-scale prohibited items detection model. EAI Endorsed Transactions on AI and Robotics. 2025;4. doi:10.4108/airo.10277

[27] Musa W, Katili MR, Ridwan W. Optimizing machine learning architectures for time series forecasting: a hybrid rvGA-eNM approach. EAI Endorsed Transactions on AI and Robotics. 2025;4. doi:10.4108/airo.9976