Feature-Context Consistency: Unsupervised Adversarial Detection with Drift Stability on Attributed Graphs

Firas Bayram
SnT, University of Luxembourg
Luxembourg
firas.bayram@uni.lu

Maxime Cordy
SnT, University of Luxembourg
Luxembourg
maxime.cordy@uni.lu

Abstract

Adversarial detection on evolving attributed graphs faces two challenges: fraud labels arrive late or not at all, and natural distributional drift erodes any trained decision boundary. Modern detectors can reach strong detection rates when labels are available and the distribution is stationary; sustaining that quality without labels and under drift is the open problem. We identify *feature-context consistency* — the alignment between a node's attributes and its aggregated neighborhood attributes — as a single structural signal that addresses both. We formalize it as the **Context Boost** (CB) score within a Restricted Boltzmann Machine framework (CB-RBM) and prove two complementary worst-case guarantees: a strict separation between the expected CB of legitimate and adversarial nodes, with gap growing linearly in model capacity, whenever the adversary's context stays far enough from legitimate neighborhood context; and a Wasserstein-Lipschitz bound on how far the CB distribution moves under natural drift. We validate on XBlock Ethereum phishing ($\sim$153K nodes) and Reddit banned-user detection ($\sim$11K nodes), with two further datasets in the appendix spanning diverse domains. With zero labels, CB-RBM achieves AUROC ≥ 0.96 on four adversarial attacks and keeps FPR close to its calibrated 5% target under drift — outperforming supervised GNNs (hundreds of labels), unsupervised graph-anomaly baselines, and a standard RBM whose FPR collapses to 89–100%. What typically requires separate mechanisms — label-free operation and drift stability — here follows from a single structural regularity of the data. Code: https://github.com/serval-uni-lu/CB-RBM.

1 Introduction

Within networked systems, adversarial actors (e.g., scammers in cryptocurrency transaction graphs, banned users on social platforms, fake reviewers in e-commerce networks) operate at scale to bypass systems' security and business protection measures. The stakes are substantial: cryptocurrency fraud alone reached $9.3 billion in 2024 [1, 2], with comparable damage from coordinated review rings, account abuse, and platform manipulation [3]. These systems are naturally represented as *attributed graphs*, where nodes carry features and edges encode relationships, and detection requires exploiting both channels together. Adversarial manipulation then consists of introducing new nodes and edges in the graph following a specific (adversarial) feature distribution.

Yet deploying detectors on these graphs faces two compounding constraints. First, *label scarcity*: adversarial entities typically operate for days or weeks before being flagged—if they are ever reported at all—leaving no reliable supervisory signal during the critical early window when the most damage is done [4]. Second, *distributional drift* [5, 6]: user behavior, network topology, and feature distributions evolve continuously through organic growth and churn [7], inducing natural

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

concept drift [8]. A detector trained on a snapshot implicitly encodes the feature distribution of that moment; as legitimate entities migrate to new regions of feature space, the learned decision boundary becomes stale—legitimate activity that was once well within the normal range now falls outside it, triggering escalating false alarms [9]. The standard remedy is periodic retraining, but this loops back to the first problem. Retraining requires fresh labels, which arrive late and incomplete. In particular, supervised fraud GNNs [10–12] depend on exactly these labels, and unsupervised graph detectors such as DOMINANT [13] and CoLA [14] carry no guarantee against drift-induced false alarms. Without a mechanism that is inherently stable under drift, practitioners are trapped in a cycle where the model is either stale or starved of supervision.

We identify *feature-context consistency*—the alignment between an entity's own attributes and the aggregated attributes of its neighbors—as a single structural signal that simultaneously addresses both problems. A legitimate entity develops feature patterns that are statistically coherent with those of its neighbors, because repeated interaction shapes both parties. An adversarial entity, by contrast, is injected without this history: its features may mimic legitimate patterns, but its neighborhood context is absent, random, or structurally inconsistent—and the adversary cannot unilaterally control the features of its neighbors. This enables *label-free detection*. The same signal is also robust to drift, because natural drift moves node features and their mean-pooled context in tandem and preserves their alignment, which gives *drift stability* without retraining. Both properties follow from the same structural fact: in attributed graphs with meaningful relationships, feature-context correlation is nonzero for legitimate entities and broken for early-stage adversarial injections.

Operationalizing this insight requires a label-free model whose score function is analytically decomposable into feature and feature-context contributions, so the loss of coherent context for adversarial nodes is expressible in closed form. Energy-based models satisfy the label-free requirement, learning an unnormalized density from legitimate data alone [15]. Among these, the Restricted Boltzmann Machine [16], a well-studied neural-network EBM, additionally provides a closed-form free energy whose per-hidden-unit decomposition supports the separation and drift bounds in §5 — unlike GNNs or alternative EBMs (score-based, energy-GAN). RBMs carry a track record in unsupervised fraud and anomaly detection [17–19], yet remain unexplored for attributed-graph adversarial detection. A standard RBM, however, models features in isolation — capturing what an entity *looks like*, not whether that appearance is *consistent with its surroundings* [20]. We extend the Gaussian-Bernoulli RBM with a gated context pathway that conditions the energy on aggregated neighborhood features, yielding the **Context-Boosted RBM (CB-RBM)**. The **Context Boost** (CB) score — the reduction in free energy from conditioning on neighborhood context — directly measures feature-context consistency and underlies both our detection mechanism and drift guarantees.

Our main contributions are as follows.

- **Method** (§4). We propose **CB-RBM**, an RBM with a GLU-gated context pathway conditioning the free energy on aggregated neighborhood features, and the **Context Boost** (CB) score—the free-energy reduction from context—as a label-free detection signal.
- **Theoretical framework** (§5). Under five empirically verifiable assumptions, we prove a strict separation between the expected CB of legitimate and adversarial nodes (Theorem 1) and a Wasserstein-Lipschitz bound on how far the CB distribution moves under drift (Theorem 2, Corollary 2)—coupling label-free detection and drift stability through a single feature-context correlation structure.
- **Detection boundary analysis** (§6.2, §6.5). The ρ_1^{eff} analysis reveals that CB detects *absence of legitimate network context* rather than feature deviation, characterizing the detection signal and delineating its scope.
- **Empirical validation** (§6). On XBlock Ethereum phishing [21] and Reddit banned users [22], CB-RBM (0 labels) achieves AUROC ≥ 0.96 on all non-white-box attacks and keeps FPR close to its calibrated 5% target under 20-step drift, ahead of supervised GNNs trained on hundreds of labels, unsupervised graph-anomaly baselines and a standard RBM whose FPR rises to 89–100%.

2 Related Work

Graph anomaly detection. Supervised fraud detectors such as CARE-GNN [10], GraphConsis [11], PC-GNN [12] and BWGNN [23] learn from labeled adversaries, which arrive late in our setting. Unsupervised detectors exploit graph structure without labels. DOMINANT [13] and Anomaly-

DAE [24] score attribute and structure reconstruction errors, GAD-NR [25] reconstructs each node's neighborhood, CoLA [14] contrasts each node with its local subgraph, CONAD [26] adds contrastive learning over augmented anomalies, and OCGNN [27] fits a one-class boundary around normal embeddings. CB-RBM uses no reconstruction target or contrastive loss. Context enters the energy of a trained model through a gated pathway, and the score is the free-energy reduction that context produces, in closed form and decomposable over hidden units (Definition 1).

Distribution shift and dynamic graphs. Concept-drift methods detect a shift and then adapt or retrain [8, 28], and change detection extends to graph streams [9]. GDN [29] uses labels to counter structural distribution shift between training and test nodes, and WhENDS [30] whitens normal edge embeddings over time to track normality shift in dynamic graphs. Dynamic-graph detectors such as StrGNN [31] and TADDY [32] model temporal evolution to score anomalous edges [33]. These approaches address shift through adaptation, robustness objectives, embedding alignment or explicit temporal modeling. In our main drift evaluation, CB-RBM keeps its trained parameters and threshold frozen, and Theorem 2 bounds how far its scores move under smooth drift.

Generalist detectors. ARC [34], UNPrompt [35] and AnomalyGFM [36] train one detector on some graphs and apply it to unseen graphs with few or no target labels. UNPrompt predicts latent node attributes from neighborhoods and AnomalyGFM uses node-neighbor representation residuals. Their shift is across graphs, whereas ours is temporal drift within one deployed graph.

Positioning. Our contribution relative to these families is the Context Boost differential and the analysis it admits, a conditional separation of expected scores (Theorem 1) and a Wasserstein bound on how far the scores move under drift (Theorem 2). Section 6.4 and Appendix D compare representatives of the node-level families under our protocol.

3 Problem Setup

Graph. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an attributed directed graph with node features $\mathbf{x}_i \in \mathbb{R}^{n_v}$ for each $i \in \mathcal{V}$. Each node represents an entity (e.g., wallet, user, reviewer) and each directed edge an interaction. Our detector uses no adversarial labels. It is trained one-class on a pool of nodes assumed legitimate, in practice accounts with no abuse report after an aging window, and its threshold is set on a held-out slice of the same pool.

Neighborhood context. The context of node i aggregates the features of its neighbors $\mathcal{N}(i)$:

$$\mathbf{c}_i = \mathbf{s} \odot \frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} \mathbf{x}_j \in \mathbb{R}^{n_v}, \tag{1}$$

where $\mathbf{s} \in \{-1, +1\}^{n_v}$ is a per-dimension sign vector that corrects for anti-homophilic dimensions, ensuring that the signed context is positively correlated with node features for legitimate nodes. The construction of $\mathbf{s}$ is detailed in §4. The mean-pool aggregator is ablated in Appendix G.

Threat model. An adversary aiming to evade detection injects a new node i that it controls (features and edges), but not the existing nodes around it; the chosen feature distribution and neighborhood structure depend on the attack strategy. We distinguish four attack strategies of increasing sophistication:

1. *Isolated Injection* — adversarial features $\mathbf{x}' \sim \mathcal{P}_{\text{fraud}}$ with context $\mathbf{c}' \sim \mathcal{Q}_{\text{rand}}$, here $\mathcal{N}(\mathbf{0}, \mathbf{I})$, drawn independently of $\mathbf{x}'$, so the node has no legitimate neighborhood context.
2. *Feature Camouflage* — legitimate features $\mathbf{x}' \sim \mathcal{P}_{\text{legit}}$ with $\mathbf{c}' \sim \mathcal{Q}_{\text{rand}}$ as above, so the node still lacks genuine context.
3. *Sybil attack* — $\mathbf{x}' \sim \mathcal{P}_{\text{fraud}}$, and the adversary links to a fully connected clique S of 50 decoys with near-legitimate features $\mathbf{x}_s = \mathbf{x}_{\ell_s} + 0.05\,\boldsymbol{\epsilon}_s$ (legitimate ℓ_s, $\boldsymbol{\epsilon}_s \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$), so its fabricated context $\mathbf{c}'$ is the mean-pool (1) over $\mathcal{N}(i) \cup S$, where $\mathcal{N}(i)$ is the neighborhood before the attack.
4. *Relation Camouflage* — $\mathbf{x}' \sim \mathcal{P}_{\text{fraud}}$, and an adaptive adversary links to the 10 highest-degree legitimate nodes H, so its partially genuine context $\mathbf{c}'$ is the mean-pool (1) over $\mathcal{N}(i) \cup H$.

These four patterns span the canonical injected-adversary threats established in graph fraud detection and adversarial graph learning [37, 38]; attack families requiring training-time graph or label manipulation (e.g., topology poisoning [39], backdoor attacks) lie outside this scope. We separately evaluate

robustness to adversarial contamination of the training pool assumed legitimate in §6.5. White-box PGD [40], which needs gradient access to the detector, is a stress test in Appendix H.

4 CB-RBM

We extend the Gaussian-Bernoulli RBM [41], a Gaussian-visible, Bernoulli-hidden variant suited to real-valued features, with a context pathway that conditions hidden unit activations on neighborhood features. The model's visible inputs are a node's features $\mathbf{v} \in \mathbb{R}^{n_v}$ and neighborhood context $\mathbf{c} \in \mathbb{R}^{n_v}$ (Eq. (1)), with binary hidden units $\mathbf{h} \in \{0,1\}^{n_h}$ as the latent layer.

Sign vector construction. The sign vector $\mathbf{s}$ in Eq. (1) is estimated from training data as $s_d = \mathrm{sign}(\mathrm{Pearson}(x_{i,d}, \bar{x}_{\mathcal{N}(i),d}))$, flipping anti-homophilic dimensions so that the signed context is positively correlated with node features across all dimensions. This is fitted once before training from feature statistics alone (no labels) and kept fixed. Because it needs no labels, it can be re-estimated on any new window of unlabeled nodes exactly as it was first fitted, and this refresh restores calibration after sign corruption (Appendix E.2).

Energy function. The joint energy is

$$E(\mathbf{v}, \mathbf{h}, \mathbf{c}; \theta) = \tfrac{1}{2}\|\mathbf{v} - \mathbf{b}_v\|^2 - \sum_{j=1}^{n_h} h_j\big(\mathbf{w}_j^\top \mathbf{v} + \hat{\mathbf{c}}_j + b_j\big), \tag{2}$$

where $\hat{\mathbf{c}}_j = (\mathbf{U}_j^\top \mathbf{c}) \cdot \sigma(\mathrm{Gate}_j^\top \mathbf{c})$ is a Gated Linear Unit [42] context embedding and $\theta = \{\mathbf{W}, \mathbf{U}, \mathrm{Gate}, \mathbf{b}_v, \mathbf{b}_h\}$. When context is absent ($\mathbf{c} = 0$), the multiplicative $\mathbf{U}_j^\top \mathbf{c}$ factor vanishes and the model recovers a standard RBM. The free energy marginalizing over $\mathbf{h}$ is

$$F(\mathbf{v}, \mathbf{c}; \theta) = \tfrac{1}{2}\|\mathbf{v} - \mathbf{b}_v\|^2 - \sum_{j=1}^{n_h} \mathrm{softplus}\big(\mathbf{w}_j^\top \mathbf{v} + \hat{\mathbf{c}}_j + b_j\big). \tag{3}$$

Context Boost score.

Definition 1 (Context Boost)**.** *The Context Boost score is*

$$\mathrm{CB}(\mathbf{v}, \mathbf{c}; \theta) = \sum_{j=1}^{n_h} \Big[\mathrm{softplus}\big(\mathbf{w}_j^\top \mathbf{v} + \hat{\mathbf{c}}_j + b_j\big) - \mathrm{softplus}\big(\mathbf{w}_j^\top \mathbf{v} + b_j\big)\Big]. \tag{4}$$

CB measures how much neighborhood context *reduces* free energy beyond features alone. The quadratic term $\frac{1}{2}|\mathbf{v} - \mathbf{b}_v|^2$ cancels exactly in the subtraction, so CB depends only on hidden-unit activations, not on feature reconstruction cost. For a legitimate node with consistent context, the context pathway activates strongly and CB is large. For an adversarial node with absent or random context, the context contribution vanishes (algebraically when $\mathbf{c}$=0, suppressed by the GLU gate when noisy) and CB is near zero (Theorem 1; Figure 1).

Training. We train exclusively on legitimate node pairs $(\mathbf{x}_i, \mathbf{c}_i) \sim \mathcal{P}_{\text{legit}}$ via Contrastive Divergence CD-1 [16]. No attack labels are used. We set the sparsity penalty $\lambda_{\text{sp}} = 0$; nonzero sparsity drives mean hidden activations below 0.5 and weakens the context contribution.

Detection. The detection score for node i is $s_i = -\mathrm{CB}(\mathbf{x}_i, \mathbf{c}_i)$. The threshold τ^* is set at the 95th percentile of $-$CB over a held-out calibration slice of the pool, setting a nominal 5% calibration target. Nodes with $s_i \geq \tau^*$ are flagged as adversarial. Figure 4 and Algorithm 1 (Appendix A) summarize the full pipeline.

Complexity. Training costs $O(N \cdot n_h \cdot n_v)$ per epoch on N training nodes and inference $O(n_h \cdot n_v)$ per node, as for a standard RBM, since the single-hop context is precomputed once and only closed-form free energies are evaluated [43].

5 Theoretical Guarantees

We establish two complementary results for the CB score. The first gives a sufficient condition under which the expected CB of legitimate nodes exceeds that of adversarial nodes. The second shows that CB remains stable under natural concept drift, explaining why a single threshold serves both detection and drift resilience. Both are worst-case certificates rather than predictions of the observed error, which Section 6 measures directly. Full proofs are deferred to Appendix B. Remarks below provide intuition and justification; constants are verified in Table 2.

5.1 Assumptions

Assumption 1 (Signed Structural Correlation)**.** *The signed mean-pool context* (1) *satisfies*

$$\mathbb{E}_{i\sim\mathcal{P}_{\text{legit}}}\big[\cos(\mathbf{x}_i,\mathbf{c}_i)\big] \geq \rho_0 > 0, \quad \textit{for some } \rho_0 \in (0,1]. \tag{5}$$

Remark 5.1. $\mathbf{s}$ corrects anti-homophilic dimensions [44] via $s_d = \text{sign}(\text{Pearson}(x_{i,d}, \bar{x}_{\mathcal{N}(i),d}))$, computed once on the training set. Per-dimension covariances are non-negative by construction, so $\rho_0 > 0$ whenever any dimension has nonzero neighbor covariance, which we measure directly (§6).

Assumption 2 (Coherent Activation)**.** *After training, a fraction $\beta_0 \in (0,1)$ of hidden units j satisfy*

$$\mathbb{P}_{(\mathbf{v},\mathbf{c})\sim\mathcal{P}_{\text{legit}}}\big[\alpha_j^{\text{feat}} \geq 0 \textit{ and } \alpha_j^{\text{ctx}} \geq 0\big] \;\geq\; p_0 > 0, \tag{6}$$

where $\alpha_j^{\text{feat}} = \mathbf{w}_j^\top \mathbf{v} + b_j$ and $\alpha_j^{\text{ctx}} = \hat{\mathbf{c}}_j$.

Remark 5.2. A nontrivial fraction of hidden units jointly activate on matched $(\mathbf{x},\mathbf{c})$ pairs. Under the Gaussian-projection and symmetry assumptions of the arcsine argument in Appendix B, $p_0 \geq \frac{1}{4} + \rho_0/(2\pi)$, which predicts $p_0 \geq 0.30$ on XBlock and 0.36 on Reddit before training. After training, $\beta_0 = 0.38$ and 0.95, that is, 97 of 256 and 122 of 128 hidden units (Table 2).

Assumption 3 (Bounded Moments)**.** $\mathbb{E}[\|\mathbf{x}\|_2^2] \leq M_x^2$ *and* $\mathbb{E}[\|\mathbf{c}\|_2^2] \leq M_c^2$.

Assumption 4 (Adversarial Context Gap)**.** *The adversarial context distribution satisfies, for every hidden unit j,*

$$\mathbb{E}_{\mathbf{c}'\sim\mathcal{Q}_{\text{rand}}}\big[\max(\hat{\mathbf{c}}'_j, 0)\big] \;\leq\; \rho_1 \,\|\mathbf{U}_j\|\, M_c, \quad \textit{for some } \rho_1 \in [0,\rho_0). \tag{7}$$

Remark 5.3. Here ρ_1 bounds positive gated-context activation under adversarial context, and $\Delta\rho = \rho_0 - \rho_1 > 0$ denotes the *detection gap*. Theorem 1 certifies separation when the adversarial positive-context bound lies sufficiently below the dominance margin of Assumption 5.

Assumption 5 (Coherent Dominance)**.** *The expected gain $G_{\mathcal{C}}$ of the coherent units of Assumption 2 on their joint-positive events exceeds the expected negative contribution N of all units by a margin $\gamma > 0$ independent of n_h, $G_{\mathcal{C}} - N \geq \gamma$ (both normalized by n_h and defined in Appendix B.2).*

Remark 5.4. Assumption 1 captures homophily [45]. Assumption 2 formalizes the consequences of training on correlated data. Assumption 3 imposes standard boundedness conditions. Assumption 4 bounds the positive gated-context activation that adversarial context can produce—a realistic constraint, since attackers cannot fabricate authentic interaction history at scale. Assumption 5 involves only the legitimate distribution and the trained model, so it is checked without adversarial data. Before training, the arcsine bound on p_0 checks the coherence of Assumption 2 on which the gain $G_{\mathcal{C}}$ rests.

5.2 Separation and Stability

Let $\Psi = \mathbb{E}[\text{CB} \mid \mathcal{P}_{\text{legit}}] > 0$ denote the expected CB under the legitimate distribution, $\Phi(\rho_1) = \mathbb{E}[\text{CB} \mid \mathcal{P}_{\text{fraud}}, \mathcal{Q}_{\text{rand}}]$ the expected CB under adversarial features and context, and $d(\mathbf{x}', \mathcal{P}_{\text{legit}}) = \inf_{\mathbf{x}\in\mathcal{P}_{\text{legit}}} \|\mathbf{x}' - \mathbf{x}\|$ the feature deviation of an adversarial node from the legitimate support.

Theorem 1 (Separation)**.** *Under Assumptions 1–5, $\Psi \geq n_h\gamma$. If, in addition, $\rho_1 \bar{u} M_c \leq \gamma - \varepsilon$ for some $\varepsilon > 0$ independent of n_h, where $\bar{u} = \frac{1}{n_h}\sum_j \|\mathbf{U}_j\|$, then $\Psi - \Phi \geq n_h\varepsilon$, a strict separation growing linearly in n_h. For any threshold $\tau^* \in (\Phi, \Psi)$, Chebyshev's inequality gives*

$$\mathbb{P}_{(\mathbf{x},\mathbf{c})\sim\mathcal{P}_{\text{legit}}}[\text{CB}(\mathbf{x},\mathbf{c};\theta^*) < \tau^*] \leq \frac{\text{Var}[\text{CB} \mid \mathcal{P}_{\text{legit}}]}{(\Psi - \tau^*)^2}, \tag{8}$$

$$\mathbb{P}_{\mathbf{x}'\sim\mathcal{P}_{\text{fraud}},\,\mathbf{c}'\sim\mathcal{Q}_{\text{rand}}}[\text{CB}(\mathbf{x}',\mathbf{c}';\theta^*) \geq \tau^*] \leq \frac{\text{Var}[\text{CB} \mid \mathcal{P}_{\text{fraud}}, \mathcal{Q}_{\text{rand}}]}{(\tau^* - \Phi)^2}. \tag{9}$$

Both bounds are finite since $\text{Var}[\text{CB}] \leq (\sum_j M_j)^2 < \infty$ (Lemma 2).

Proof sketch. Step 1: Decomposition. Write $\mathrm{CB}(\mathbf{v}, \mathbf{c}) = \sum_j \Delta_j$ where $\Delta_j = \mathrm{softplus}(\alpha_j^{\mathrm{feat}} + \alpha_j^{\mathrm{ctx}}) - \mathrm{softplus}(\alpha_j^{\mathrm{feat}})$.

Step 2: Legitimate lower bound. On the joint-positive event (Assumption 2), Lemma 1 gives $\Delta_j \geq \mathrm{softplus}(\alpha_j^{\mathrm{ctx}}) - \log 2 \geq 0$. Summing the coherent units' gains and subtracting every unit's expected negative contribution, Assumption 5 gives $\Psi \geq n_h(G_{\mathcal{C}} - N) \geq n_h\gamma > 0$.

Step 3: Adversarial upper bound. Whatever the features, $\Delta_j \leq \max(\hat{\mathbf{c}}_j, 0)$, so Assumption 4 gives $\Phi \leq n_h\rho_1\bar{u}M_c$ and hence $\Psi - \Phi \geq n_h\varepsilon$.

Step 4: Tail bounds. Since $\mathrm{Var}[\mathrm{CB}] \leq (\sum_j M_j)^2$ (Lemma 2), both variances are finite and Chebyshev yields the bounds in (8)–(9). See Appendix B.2 for the full proof. □

Corollary 1 (Detection Guarantee)**.** *Under the conditions of Theorem 1, setting $\tau^* = (\Psi + \Phi)/2$ gives*

$$\mathrm{FPR} \;\leq\; \frac{4\,\mathrm{Var}[\mathrm{CB} \mid \mathcal{P}_{\mathrm{legit}}]}{(\Psi - \Phi)^2}, \qquad 1 - \mathrm{TPR} \;\leq\; \frac{4\,\mathrm{Var}[\mathrm{CB} \mid \mathcal{P}_{\mathrm{fraud}}, \mathcal{Q}_{\mathrm{rand}}]}{(\Psi - \Phi)^2}.$$

Remark 5.5. Separation is certified when the adversarial positive-context bound $\rho_1\bar{u}M_c$ lies below the legitimate dominance margin γ, and the certified margin shrinks as ρ_1 grows. The Chebyshev bounds are finite but can be vacuous, since $(\sum_j M_j)^2$ exceeds the measured $\mathrm{Var}[\mathrm{CB}]$ by $8{,}300\times$ on XBlock and $237{,}000\times$ on Reddit, so detection is measured on the full score distributions (Figure 1, Table 1).

Theorem 2 (Drift Stability)**.** *Let $\{(\mathbf{x}_t, \mathbf{c}_t) \sim \mathcal{P}_t\}_{t=0}^{T}$ undergo natural drift where:*

(i) Correlation structure persists: $\mathbb{E}[\cos(\mathbf{x}_t, \mathbf{c}_t)] \geq \rho_0$ for all t.

(ii) Drift is smooth in Wasserstein-1, $W_1(\mathcal{P}_t, \mathcal{P}_{t-1}) \leq \delta_{\mathrm{drift}}$ for all t, and every $\mathcal{P}_t$ is supported on $\{\|\mathbf{c}\|_2 \leq M_c\}$.

Then for CB-RBM trained on $\mathcal{P}_0$,

$$\left|\mathbb{E}[\mathrm{CB}(\mathbf{x}_t, \mathbf{c}_t)] - \mathbb{E}[\mathrm{CB}(\mathbf{x}_0, \mathbf{c}_0)]\right| \leq L\,t\,\delta_{\mathrm{drift}}, \tag{10}$$

where $L = L_v + L_c$ is the Lipschitz constant of CB on that region (Lemma 3). Corollary 3 gives a worst-case horizon over which the legitimate mean remains above the midpoint threshold $(\Psi + \Phi)/2$.

Proof sketch. CB is L-Lipschitz (Lemma 3); Kantorovich–Rubinstein gives $|\mathbb{E}[\mathrm{CB}_t] - \mathbb{E}[\mathrm{CB}_{t-1}]| \leq L \cdot W_1(\mathcal{P}_t, \mathcal{P}_{t-1}) \leq L\delta_{\mathrm{drift}}$. Telescoping over t steps yields (10); full proofs, including Corollary 2 and the drift-horizon corollary, in Appendix B. □

Corollary 2 (Quantile Stability)**.** *Under condition (ii) of Theorem 2, the quantile functions Q_t of the detection score $-\mathrm{CB}(\mathbf{x}_t, \mathbf{c}_t)$ satisfy $\int_0^1 |Q_t(u) - Q_0(u)|\,du \leq L\,t\,\delta_{\mathrm{drift}}$.*

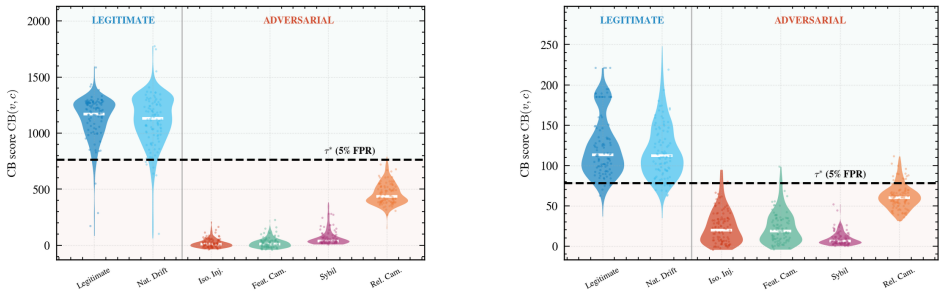
Remark 5.6. Natural drift preserves feature-context correlation and causes CB to change gradually at rate $L\,\delta_{\mathrm{drift}}$ per step. Adversarial injection breaks this correlation and causes immediate CB suppression. This asymmetry is what enables a single threshold to simultaneously detect adversaries and tolerate drift. Corollary 2 controls the score quantiles on average but not the 95th percentile pointwise, which Appendix I.2 measures directly.

General attacks. The separation condition depends on an attack only through the positive-context bound ρ_1 of Assumption 4, since $\Phi \leq n_h\rho_1\bar{u}M_c$ holds for any adversarial features. Feature deviation acts as a bounded perturbation. At fixed context, it changes each unit's contribution by at most $\|\mathbf{w}_j\|_2\,d$ (Appendix B.2). The context-interpolation sweep in Appendix J moves the attack context from its revealing to its camouflaging end. Detection falls below 50% once the camouflaging share exceeds $\alpha^* \approx 0.56$ on XBlock and 0.58 on Reddit, and the hardest camouflage is whichever context brings ρ_1^{eff} closest to ρ_0.

6 Experiments

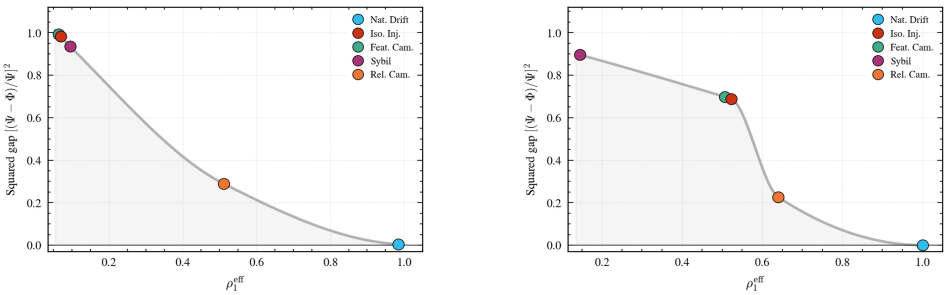
6.1 Setup

Datasets. We use two attributed graphs from distinct domains: **(1) XBlock** (153K Ethereum wallet nodes, 0.7% phishing, 38 features, $\rho_0 = 0.30$) and **(2) Reddit** (11K social-media users, 3.3% banned,



(a) XBlock (b) Reddit

Figure 1: CB score distributions, with calibrated threshold τ^* (5% FPR).



(a) XBlock ($\rho_0 = 0.30$) (b) Reddit ($\rho_0 = 0.68$)

Figure 2: Squared separation gap $[(\Psi-\Phi)/\Psi]^2$ vs. ρ_1^{eff} (Thm. 1).

73 features, $\rho_0 = 0.68$). Context vectors are a per-dimension sign-corrected mean-pool over graph neighbors (Eq. (1)), preserving $\rho_0 > 0$ regardless of homophily direction. Full dataset details and verified Assumption 1–3 constants in Appendix C; two further datasets — Tolokers (crowdsourcing) and YelpChi (e-commerce review fraud) — evaluated in Appendix K.

Baselines. We benchmark CB-RBM against six baselines from the GADBench [46] suite. GraphSAGE [47], GAT [48] and GCN [49] represent supervised fraud GNNs, foundational variants that GADBench shows match or exceed specialized fraud-GNN refinements, and DOMINANT [13], CoLA [14] and GAE [50] represent unsupervised attributed-graph anomaly detection. A standard RBM without context (Std-RBM) is the ablation. We also compare against nine detectors from the neighborhood-aware and drift-aware classes closest to our setting. These are the supervised XGBGraph [46], BWGNN [23] and Graph Transformer [51], the shift-robust GDN [29], the generalist ARC [34], the reconstruction-based GAD-NR [25] and AnomalyDAE [24], the contrastive CONAD [26] and the one-class OCGNN [27]. Two further ablations, a parameter-free signed-cosine score and a context-conditioned MLP without the RBM and GLU (Context-MLP), isolate our formulation. The nine detectors follow their published configurations, with model selection on a held-out validation split (Appendix D). Supervised GNNs use 932 / 166 labeled adversaries on XBlock / Reddit; all other methods use **zero labels**.

Metrics. *TPR@5%FPR*: true positive rate at a 5%-FPR-calibrated threshold on legitimate nodes (no adversarial labels needed). *Stability*: $1 - \text{FPR}_{\text{drift}}$, fraction of legitimate nodes correctly classified under drift. *Combined Score*: $\text{TPR@5\%FPR} \times \text{Stability}$ per attack; we also report the mean over the four attacks as "average Combined".

CB-RBM configuration. n_h=256 for XBlock and n_h=128 for Reddit (capacity ablation in Appendix G), 100 epochs of CD-1, sparsity cost = 0.

Table 1: Attack benchmark. Each cell: AUROC / TPR@5%FPR (%). Best per row in **bold**.

	Attack	Energy-based		Supervised			Unsupervised		
		CB-RBM	Std-RBM	SAGE	GAT	GCN	DOM.	CoLA	GAE
XBlock	Isolated Inj.	**99.9/100**	86.8/51.6	97.6/90.1	91.7/75.2	95.4/80.8	96.9/78.1	56.4/17.2	98.1/89.5
	Feat. Camo	**99.9/100**	51.3/11.0	92.6/77.8	54.6/10.2	84.0/29.7	96.4/72.5	57.3/16.2	96.7/78.7
	Sybil	**99.8/100**	86.8/51.6	98.8/94.9	99.4/**100**	99.6/**100**	98.1/98.8	62.5/21.6	96.6/77.4
	Rel. Camo	97.6/96.8	86.8/51.6	**98.6**/92.4	98.2/**97.8**	97.8/87.8	97.0/86.9	72.1/20.2	97.7/88.9
Reddit	Isolated Inj.	**99.4**/97.2	60.0/10.7	61.9/0.5	73.5/38.3	82.2/62.8	98.0/91.3	93.7/83.8	99.1/**97.3**
	Feat. Camo	**99.6/98.2**	52.4/4.3	55.1/0.0	58.4/18.2	57.3/29.5	90.6/79.7	92.7/79.0	90.5/86.5
	Sybil	**100/100**	60.0/10.7	68.2/1.8	92.9/66.7	81.8/33.3	99.9/**100**	85.8/43.2	97.4/86.8
	Rel. Camo	**97.3/88.2**	60.0/10.7	83.2/19.7	82.1/5.2	83.7/19.2	93.2/52.8	77.9/18.7	89.3/48.5

6.2 Separation Gap and Theory Validation

A clear structural divide separates legitimate from adversarial CB score distributions on both datasets (Fig. 1): legitimate and naturally drifted nodes concentrate well above the threshold τ^*, while all four attack types cluster predominantly below it — regardless of whether the adversary copies features, fabricates context, or rewires edges. On each dataset, a single threshold calibrated from the legitimate distribution generalizes across all four attacks.

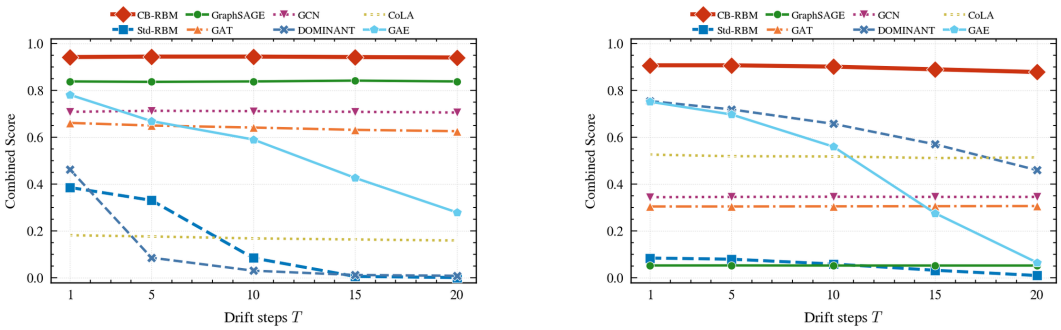
To understand why some attacks come closer to the threshold than others, we measure ρ_1^{eff}, the fraction of the RBM's learned legitimate-context activation that each attack achieves (0 = no legitimate context, 1 = indistinguishable from legitimate) — the effective adversarial correlation through the trained GLU weights, which rises when an adversary connects to legitimate hubs — and plot it against the squared separation gap (Figure 2; the squared form reflects the Chebyshev error bound of Theorem 1). The result traces a smooth, monotonically decreasing curve on both datasets, consistent with Theorem 1, whose certified margin shrinks as ρ_1 grows. Sybil's fabricated context is rejected by the learned weights; Relation Camouflage, the only attack that acquires connections to real legitimate hubs, reaches the highest ρ_1^{eff} (0.51–0.64) and the steepest gap reduction. Natural Drift preserves the legitimate correlation structure ($\rho_1^{\text{eff}} \approx 1$) and sits at near-zero gap — explaining why the same threshold that catches adversaries does not trigger false alarms under drift.

6.3 Attack Benchmark

Table 1 compares per-attack detection. The four attacks span a spectrum of adversarial context quality, from uncorrelated (Isolated Injection, Feature Camouflage) through fabricated (Sybil) to partially genuine (Relation Camouflage). CB-RBM's AUROC tracks this spectrum cleanly on both datasets: it stays above 99% on three attacks and drops to $\sim 97\%$ on Relation Camouflage (where the adversary acquires genuine hub connections), consistent with the $\Delta\rho$-dependent separation of Theorem 1. Across all four attacks on both datasets ($14\times$ scale, $\rho_0 = 0.30$ vs 0.68), CB-RBM maintains AUROC $\geq 96\%$ and TPR@5%FPR $\geq 88\%$ at zero labels. Over 3 seeds, the standard deviation of CB-RBM's AUROC is at most 0.1 points on XBlock and 1.2 on Reddit, and that of its TPR at most 0.3 and 4.5 points. Supervised GNNs, despite their full label budget, remain undertrained in this regime (best Reddit supervised result: GAT at 67% TPR on Sybil). Unsupervised baselines avoid labels but their signal does not transfer reliably across datasets (CoLA: $84\% \to 17\%$ TPR).

6.4 Drift Stability

Following standard practice in the concept drift literature [8, 28], we model gradual drift as an incremental random walk in feature space over T steps, with neighborhood context recomputed at each step to reflect the evolving features. Figure 3 reports the Combined Score (detection $\times$ stability, §6.1) as drift accumulates, using each method's mean TPR across the four attacks of Table 1 ($\text{FPR}_{\text{drift}}$ is measured on legitimate nodes only, so we average TPR across attacks but use a single FPR curve); additional drift regimes and real temporal validation are in Appendix I.



(a) XBlock (b) Reddit

Figure 3: Combined Score under gradual drift across 20 time steps.

Figure 3 reveals three patterns. First, reconstruction-based detectors (GAE, DOMINANT), strong on the static benchmark (§6.3), collapse as their learned reconstruction targets become stale — DOMINANT drops to near zero on XBlock, and GAE from 0.75 to 0.07 on Reddit. Second, supervised GNNs split by label availability. On XBlock, where labels provide a strong training signal, GNNs learn robust decision boundaries that generalize under drift, achieving Combined Scores up to 0.84 (GraphSAGE). On Reddit, where labels are scarce, the same architectures remain undertrained (§6.3), and their Combined Score remains below 0.36 throughout — stable only because there is weak TPR to degrade. Third, CB-RBM is the only method that consistently leads on both datasets, maintaining average Combined ≥ 0.87 at $T{=}20$ in both the label-rich and label-scarce regimes. Theorem 2 bounds how far the scores move under smooth drift, and Figure 3 measures the resulting detection and false alarms at the frozen threshold. Across five chronological XBlock windows from 2017 through 2018, with the model, sign vector and threshold frozen on the earliest window, CB-RBM keeps the legitimate-wallet FPR at or below 5%, reaching 3.1% by late 2018, and detects 100% of isolated injections built from real held-out phishing features. GAT and GraphSAGE, which score the phishing wallets in the original graph, detect 84–89% and 81% (Appendix I.4).

Table 1 and Figure 3 report eight methods per attack and over the full drift trajectory. We compare the other detectors of §6.1 with the same Combined Score at the last drift step, $T{=}20$ (Appendix D). At that step, the strongest baselines are GraphSAGE (0.84) and XGBGraph (0.81, with TPR 0.98 but drift FPR 17.8%) on XBlock and GAD-NR (0.63) on Reddit, against 0.94 and 0.88 for CB-RBM. Over five seeds, the Combined Scores of XGBGraph, Graph Transformer, BWGNN and GDN vary with a standard deviation below 0.05. GAD-NR misses the self-consistent Sybil clique on XBlock (0% TPR). GDN reaches 10.4% drift FPR on XBlock and ARC 47.2% on Reddit. The ablations show that each ingredient fails on its own. The RBM without context collapses under drift, the signed-cosine score reaches only 0.17 and 0.57 Combined, and Context-MLP reaches a static AUROC of at least 0.95 on every Reddit attack yet raises 96% false alarms under drift. The principle supplies detection, and the closed-form differential supplies the calibration that survives drift.

6.5 Discussion

What does CB detect? The CB score fires on *absence of legitimate neighborhood context* rather than on feature deviation itself. The GLU gate suppresses the context pathway whenever the aggregated neighborhood correlation departs from the pattern learned during training, so a node is flagged for failing to corroborate the model's learned correlation pattern, a stricter test than raw feature-context geometry. Controlled decomposition experiments (Appendix F) confirm this on both datasets: legitimate features paired with random or absent context are flagged as reliably as adversarial features with random context (Combined ≥ 0.91), while legitimate features with legitimate context consistently evade detection (Combined ≤ 0.11). An XBlock trace gives $\mathrm{CB} = 1046$ for a median legitimate node, 952 after 20 drift steps, and 28 for an injected adversary (Appendix F). The same mechanism also bounds what CB-RBM can catch — an adversary who acquires genuine network connections eventually restores the correlation signal and evades, consistent with the ρ_1^{eff} trajectory in Figure 2.

Why the drift bound is conservative. Theorem 2 bounds the worst case: it allows every feature coordinate to move adversarially within the drift budget. Natural drift does not look like this, because features and their aggregated context co-evolve through the mean-pool operator and preserve their legitimate correlation step-by-step; movement along this correlated direction lies in a low-sensitivity region of the softplus differences in Eq. (4). The observed CB shift at $T{=}20$ is therefore smaller than the bound evaluated at the measured value of M_c by more than two orders of magnitude — a $526\times$ slack on XBlock and a $4{,}540\times$ slack on Reddit (Table 12). At $T{=}20$, the empirical 95th percentile of the legitimate scores shifts by 42.7 on XBlock and 3.4 on Reddit, while mean FPR at the frozen threshold reaches 6.1% and 7.1% (Appendix I.2).

False-positive recovery under drift. When adversarial context approaches legitimate context, supervised baselines with adequate labels can detect better. YelpChi, where the raw correlation gap is negative ($\Delta\rho{=}-0.013$), tests this regime (Appendix K.2). Even there, CB-RBM's drift stability pays off as a no-cost filter stacked atop those supervised baselines — on YelpChi it eliminates 87.6–93.5% of each baseline's drift-induced false alarms with no extra labels, training, or pipeline changes (Appendix L).

Ablation. The context pathway is what CB-RBM relies on; the surrounding architectural choices are largely fungible. Removing context (reducing the model to a free-energy RBM) collapses Combined Score by two to three orders of magnitude on both datasets, whereas varying hidden capacity within each dataset's stable range or swapping the pooling operator (mean, max, or attention) leaves detection near its ceiling (Combined ≥ 0.87 at the per-dataset optima; Appendix G).

Robustness. Contaminating the training and calibration pools with up to 10% fraud flips no sign dimension and keeps FPR at or below 5.0%, although Relation Camouflage TPR on XBlock falls to 39.0% (Appendix E.1). A stale sign vector raises FPR without hurting Isolated Injection detection, and re-estimating it without labels restores the 5% target (Appendix E.2). Drift that delays, weakens or reverses the feature-context link keeps FPR at or below 6.2% on XBlock and 13.2% on Reddit at $T{=}20$, and when all eight methods are recalibrated under co-evolving drift, CB-RBM keeps the highest Combined Score (Appendix I.3). With a Sybil clique that grows from 2 to 50 members over 50 steps, CB-RBM flags 100% of the attackers on both datasets from step 10, against 51.3% on XBlock and 9.5% on Reddit for a standard RBM (Appendix I.5).

Limitations. CB-RBM relies on $\rho_0 > 0$, which can fail in random or adversarially curated topologies. Detection also degrades as $\Delta\rho$ collapses, allowing supervised baselines with adequate labels to perform better. The method is most effective against adversaries who lack legitimate neighborhood history; as the ρ_1^{eff} analysis shows, an attacker who successfully acquires genuine hub connections (Relation Camouflage) reduces the detection gap and, in the limit, becomes indistinguishable from legitimate users (Appendices J and L). Substantial topology change can impair calibration at the frozen threshold, and a label-free threshold refresh then restores the 5% target (Appendix I.3). Finally, white-box adversaries with gradient access to the detector can lower detection (Appendix H), and our drift analysis assumes smooth drift, so abrupt or adversarial distribution shifts may need mechanisms beyond the algebraic cancellation we exploit.

7 Conclusion

We presented CB-RBM, a label-free detector exploiting feature-context consistency: the same structural signal — nonzero correlation between a node's features and its neighborhood context — simultaneously enables detection (broken for injected adversaries) and keeps calibration under drift (preserved under population-level shifts). On XBlock (153K Ethereum wallets) and Reddit (11K users), CB-RBM achieves AUROC ≥ 0.96 on all non-white-box attacks at zero labels and maintains Combined Score ≥ 0.87 at $T{=}20$ under drift, while supervised GNNs underperform even with their full label set, unsupervised baselines collapse under drift, and Std-RBM loses calibration entirely. The ρ_1^{eff} analysis reveals a clear detection boundary: CB-RBM detects *absence of legitimate network context*, not feature deviation — explaining both its strengths and its fundamental limit against adversaries who acquire genuine connections. Two extensions follow naturally: dynamic graphs with streaming edges, where the same Lipschitz structure carries over; and porting the Context Boost principle to other generative architectures (Appendix M).

Acknowledgments and Disclosure of Funding

This work was funded in part by the Luxembourg National Research Fund (FNR), grant reference NCER22/IS/16570468/NCER-FT, through the FutureFinTech National Centre of Excellence in Research and Innovation, co-funded by the Luxembourg Ministry of Finance. For the purpose of open access, and in fulfilment of the obligations arising from the grant agreement, the authors have applied a Creative Commons Attribution 4.0 International (CC BY 4.0) license to any Author Accepted Manuscript version arising from this submission. This work was also supported by the Luxembourg AI Factory (L-AIF, grant agreement ID 1012343366). The authors declare no competing interests.

References

[1] Federal Bureau of Investigation. 2023 internet crime report. Technical report, Internet Crime Complaint Center (IC3), 2023. URL https://www.ic3.gov/annualreport/reports/2023_ic3report.pdf.

[2] Federal Bureau of Investigation. 2024 internet crime report. Technical report, Internet Crime Complaint Center (IC3), 2024. URL https://www.ic3.gov/AnnualReport/Reports/2024_IC3Report.pdf.

[3] Ishai Rosenberg, Asaf Shabtai, Yuval Elovici, and Lior Rokach. Adversarial machine learning attacks and defense methods in the cyber security domain. *ACM Computing Surveys (CSUR)*, 54(5):1–36, 2021.

[4] Jiajing Wu, Qi Yuan, Dan Lin, Wei You, Weili Chen, Chuan Chen, and Zibin Zheng. Who are the phishers? phishing scam detection on ethereum via network embedding. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 52(2):1156–1166, 2020.

[5] Qiao Yuan, Sheng-Uei Guan, Pin Ni, Tianlun Luo, Prudence Wong, Victor Chang, and Ka Lok Man. Continual graph learning: A survey. *Pattern Recognition*, page 113600, 2026.

[6] Man Wu, Xin Zheng, Qin Zhang, Xiao Shen, Xiong Luo, Xingquan Zhu, and Shirui Pan. Graph learning under distribution shifts: A comprehensive survey on domain adaptation, out-of-distribution, and continual learning. *arXiv preprint arXiv:2402.16374*, 2024.

[7] Raphael Auer, Bernhard Haslhofer, Stefan Kitzler, Pietro Saggese, and Friedhelm Victor. The technology of decentralized finance (defi). *Digital Finance*, 6(1):55–95, 2024.

[8] Jie Lu, Anjin Liu, Fan Dong, Feng Gu, Joao Gama, and Guangquan Zhang. Learning under concept drift: A review. *IEEE transactions on knowledge and data engineering*, 31(12): 2346–2363, 2018.

[9] Daniele Zambon, Cesare Alippi, and Lorenzo Livi. Concept drift and anomaly detection in graph streams. *IEEE transactions on neural networks and learning systems*, 29(11):5592–5605, 2018.

[10] Yingtong Dou, Zhiwei Liu, Li Sun, Yutong Deng, Hao Peng, and Philip S Yu. Enhancing graph neural network-based fraud detectors against camouflaged fraudsters. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 315–324, 2020.

[11] Zhiwei Liu, Yingtong Dou, Philip S Yu, Yutong Deng, and Hao Peng. Alleviating the inconsistency problem of applying graph neural network to fraud detection. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pages 1569–1572, 2020.

[12] Yang Liu, Xiang Ao, Zidi Qin, Jianfeng Chi, Jinghua Feng, Hao Yang, and Qing He. Pick and choose: a gnn-based imbalanced learning approach for fraud detection. In *Proceedings of the web conference 2021*, pages 3168–3177, 2021.

[13] Kaize Ding, Jundong Li, Rohit Bhanushali, and Huan Liu. Deep anomaly detection on attributed networks. In *Proceedings of the 2019 SIAM international conference on data mining*, pages 594–602. SIAM, 2019.

[14] Yixin Liu, Zhao Li, Shirui Pan, Chen Gong, Chuan Zhou, and George Karypis. Anomaly detection on attributed networks via contrastive self-supervised learning. *IEEE transactions on neural networks and learning systems*, 33(6):2378–2392, 2021.

[15] Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. *Advances in neural information processing systems*, 32, 2019.

[16] Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.

[17] Apapan Pumsirirat. Credit card fraud detection using deep learning based on auto-encoder and restricted boltzmann machine. *International Journal of advanced computer science and applications*, 2018.

[18] Ugo Fiore, Francesco Palmieri, Aniello Castiglione, and Alfredo De Santis. Network anomaly detection with the restricted boltzmann machine. *Neurocomputing*, 122:13–23, 2013.

[19] Łukasz Korycki and Bartosz Krawczyk. Adversarial concept drift detection under poisoning attacks for robust data stream mining. *Machine learning*, 112(10):4013–4048, 2023.

[20] Geoffrey E Hinton. A practical guide to training restricted boltzmann machines. In *Neural Networks: Tricks of the Trade: Second Edition*, pages 599–619. Springer, 2012.

[21] Liang Chen, Jiaying Peng, Yang Liu, Jintang Li, Fenfang Xie, and Zibin Zheng. Phishing scams detection in ethereum transaction network. *ACM Transactions on Internet Technology (TOIT)*, 21(1):1–16, 2020.

[22] Srijan Kumar, Xikun Zhang, and Jure Leskovec. Predicting dynamic embedding trajectory in temporal interaction networks. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1269–1278, 2019.

[23] Jianheng Tang, Jiajin Li, Ziqi Gao, and Jia Li. Rethinking graph neural networks for anomaly detection. In *International conference on machine learning*, pages 21076–21089. PMLR, 2022.

[24] Haoyi Fan, Fengbin Zhang, and Zuoyong Li. AnomalyDAE: Dual autoencoder for anomaly detection on attributed networks. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5685–5689. IEEE, 2020.

[25] Amit Roy, Juan Shu, Jia Li, Carl Yang, Olivier Elshocht, Jeroen Smeets, and Pan Li. GAD-NR: Graph anomaly detection via neighborhood reconstruction. In *Proceedings of the 17th ACM international conference on web search and data mining*, pages 576–585, 2024.

[26] Zhiming Xu, Xiao Huang, Yue Zhao, Yushun Dong, and Jundong Li. Contrastive attributed network anomaly detection with data augmentation. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 444–457. Springer, 2022.

[27] Xuhong Wang, Baihong Jin, Ying Du, Ping Cui, Yingshui Tan, and Yupu Yang. One-class graph neural networks for anomaly detection in attributed networks. *Neural computing and applications*, 33(18):12073–12085, 2021.

[28] João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM computing surveys (CSUR)*, 46(4):1–37, 2014.

[29] Yuan Gao, Xiang Wang, Xiangnan He, Zhenguang Liu, Huamin Feng, and Yongdong Zhang. Alleviating structural distribution shift in graph anomaly detection. In *Proceedings of the sixteenth ACM international conference on web search and data mining*, pages 357–365, 2023.

[30] Xiaoyang Xu, Xiaofeng Lin, Koh Takeuchi, Kyohei Atarashi, and Hisashi Kashima. Robust anomaly detection under normality distribution shift in dynamic graphs. *arXiv preprint arXiv:2509.17400*, 2025.

[31] Lei Cai, Zhengzhang Chen, Chen Luo, Jiaping Gui, Jingchao Ni, Ding Li, and Haifeng Chen. Structural temporal graph neural networks for anomaly detection in dynamic graphs. In *Proceedings of the 30th ACM international conference on Information & Knowledge Management*, pages 3747–3756, 2021.

[32] Yixin Liu, Shirui Pan, Yu Guang Wang, Fei Xiong, Liang Wang, Qingfeng Chen, and Vincent CS Lee. Anomaly detection in dynamic graphs via transformer. *IEEE Transactions on Knowledge and Data Engineering*, 35(12):12081–12094, 2021.

[33] Ocheme Anthony Ekle and William Eberle. Anomaly detection in dynamic graphs: A comprehensive survey. *ACM Transactions on Knowledge Discovery from Data*, 18(8):1–44, 2024.

[34] Yixin Liu, Shiyuan Li, Yu Zheng, Qingfeng Chen, Chengqi Zhang, and Shirui Pan. ARC: A generalist graph anomaly detector with in-context learning. *Advances in Neural Information Processing Systems*, 37:50772–50804, 2024.

[35] Chaoxi Niu, Hezhe Qiao, Changlu Chen, Ling Chen, and Guansong Pang. Zero-shot generalist graph anomaly detection with unified neighborhood prompts. *arXiv preprint arXiv:2410.14886*, 2024.

[36] Hezhe Qiao, Chaoxi Niu, Ling Chen, and Guansong Pang. AnomalyGFM: Graph foundation model for zero/few-shot anomaly detection. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 2326–2337, 2025.

[37] Leman Akoglu, Hanghang Tong, and Danai Koutra. Graph based anomaly detection and description: a survey. *Data mining and knowledge discovery*, 29(3):626–688, 2015.

[38] Lichao Sun, Yingtong Dou, Carl Yang, Kai Zhang, Ji Wang, Philip S Yu, Lifang He, and Bo Li. Adversarial attack and defense on graph data: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 35(8):7693–7711, 2022.

[39] Daniel Zügner, Amir Akbarnejad, and Stephan Günnemann. Adversarial attacks on neural networks for graph data. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2847–2856, 2018.

[40] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[41] Max Welling, Michal Rosen-Zvi, and Geoffrey E Hinton. Exponential family harmoniums with an application to information retrieval. *Advances in neural information processing systems*, 17, 2004.

[42] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *International conference on machine learning*, pages 933–941. PMLR, 2017.

[43] Shuangfei Zhai, Yu Cheng, Weining Lu, and Zhongfei Zhang. Deep structured energy based models for anomaly detection. In *International conference on machine learning*, pages 1100–1109. PMLR, 2016.

[44] Jiong Zhu, Yujun Yan, Lingxiao Zhao, Mark Heimann, Leman Akoglu, and Danai Koutra. Beyond homophily in graph neural networks: Current limitations and effective designs. *Advances in neural information processing systems*, 33:7793–7804, 2020.

[45] Kazi Zainab Khanam, Gautam Srivastava, and Vijay Mago. The homophily principle in social network analysis: A survey. *Multimedia Tools and Applications*, 82(6):8811–8854, 2023.

[46] Jianheng Tang, Fengrui Hua, Ziqi Gao, Peilin Zhao, and Jia Li. GADBench: Revisiting and benchmarking supervised graph anomaly detection. *Advances in Neural Information Processing Systems*, 36:29628–29653, 2023.

[47] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.

[48] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

[49] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[50] Thomas N Kipf and Max Welling. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*, 2016.

[51] Yunsheng Shi, Zhengjie Huang, Shikun Feng, Hui Zhong, Wenjin Wang, and Yu Sun. Masked label prediction: Unified message passing model for semi-supervised classification. *arXiv preprint arXiv:2009.03509*, 2020.

[52] Michel X Goemans and David P Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM (JACM)*, 42(6):1115–1145, 1995.

[53] Kay Liu, Yingtong Dou, Xueying Ding, Xiyang Hu, Ruitong Zhang, Hao Peng, Lichao Sun, and Philip S Yu. PyGOD: A python library for graph outlier detection. *Journal of Machine Learning Research*, 25(141):1–9, 2024.

[54] Bryan Hooi, Hyun Ah Song, Alex Beutel, Neil Shah, Kijung Shin, and Christos Faloutsos. FRAUDAR: Bounding graph fraud in the face of camouflage. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 895–904, 2016.

[55] Oleg Platonov, Denis Kuznedelev, Michael Diskin, Artem Babenko, and Liudmila Prokhorenkova. A critical look at the evaluation of GNNs under heterophily: Are we really making progress? *arXiv preprint arXiv:2302.11640*, 2023.

[56] Shebuti Rayana and Leman Akoglu. Collective opinion spam detection: Bridging review networks and metadata. In *Proceedings of the 21th acm sigkdd international conference on knowledge discovery and data mining*, pages 985–994, 2015.

[57] Jens Behrmann, Will Grathwohl, Ricky TQ Chen, David Duvenaud, and Jörn-Henrik Jacobsen. Invertible residual networks. In *International conference on machine learning*, pages 573–582. PMLR, 2019.

A Pipeline and Algorithm

Figure 4 summarizes CB-RBM from its inputs to the detection decision, and Algorithm 1 gives the full procedure, including the label-free refresh.

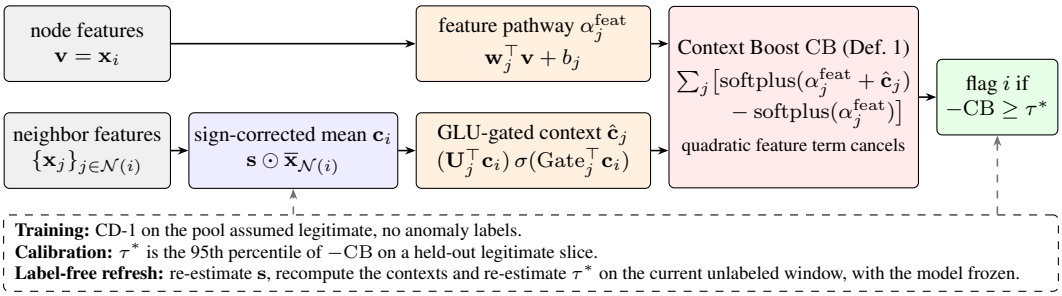


Figure 4: CB-RBM pipeline. The sign-corrected neighbor mean $\mathbf{c}_i$ enters only through the GLU-gated pathway, and the Context Boost is the free-energy reduction due to context, in which the quadratic feature term cancels. The sign vector $\mathbf{s}$ and the threshold τ^* are label-free statistics that can be refreshed on a current unlabeled window without retraining (dashed arrows, Algorithm 1).

Algorithm 1 CB-RBM: training, calibration, scoring, and label-free refresh

Input: graph $\mathcal{G}$, pool $\mathcal{V}_{\text{L}}$ assumed legitimate (no anomaly labels), calibration slice $\mathcal{V}_{\text{cal}} \subset \mathcal{V}_{\text{L}}$, capacity n_h, target FPR 5%.

1: *Sign vector.* $s_d \leftarrow \text{sign}\big(\text{Pearson}_{i \in \mathcal{V}_{\text{L}}}(x_{i,d}, \bar{x}_{\mathcal{N}(i),d})\big)$ for $d = 1, \ldots, n_v$ (features only)

2: *Context.* $\mathbf{c}_i \leftarrow \mathbf{s} \odot \frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} \mathbf{x}_j$ for all i Eq. (1)

3: *Training.* Fit $\theta = \{\mathbf{W}, \mathbf{U}, \text{Gate}, \mathbf{b}_v, \mathbf{b}_h\}$ by CD-1 on $\{(\mathbf{x}_i, \mathbf{c}_i)\}_{i \in \mathcal{V}_{\text{L}} \setminus \mathcal{V}_{\text{cal}}}$ with $\lambda_{\text{sp}} = 0$

4: *Calibration.* $\tau^* \leftarrow$ 95th percentile of $\{-\text{CB}(\mathbf{x}_i, \mathbf{c}_i)\}_{i \in \mathcal{V}_{\text{cal}}}$ Eq. (4)

5: *Scoring.* At time t, recompute $\mathbf{c}_i$ from the current neighbors of node i and flag i if $-\text{CB}(\mathbf{x}_i, \mathbf{c}_i) \geq \tau^*$, with θ frozen

6: *Label-free refresh* (optional, e.g., on a schedule or after substantial topology change). Re-estimate $\mathbf{s}$ as in line 1, recompute the contexts as in line 2 and re-estimate τ^* as in line 4 on the current unlabeled window, keeping θ frozen. No labels and no retraining.

B Full Proofs

We restate all results and provide complete proofs. Recall the CB score

$$\text{CB}(\mathbf{v}, \mathbf{c};\, \theta) \;=\; \sum_{j=1}^{n_h} \underbrace{\Big[\text{softplus}\big(\alpha_j^{\text{feat}} + \alpha_j^{\text{ctx}}\big) - \text{softplus}\big(\alpha_j^{\text{feat}}\big)\Big]}_{\Delta_j(\mathbf{v}, \mathbf{c})}, \tag{11}$$

where $\text{softplus}(x) = \log(1 + e^x)$, $\alpha_j^{\text{feat}} = \mathbf{w}_j^\top \mathbf{v} + b_j$, and $\alpha_j^{\text{ctx}} = \hat{\mathbf{c}}_j$.

B.1 Supporting Lemmas

Lemma 1 (Softplus monotone increment)**.** *For all $a, b \geq 0$,*

$$\text{softplus}(a + b) - \text{softplus}(a) \;\geq\; \text{softplus}(b) - \log 2.$$

Proof. Write

$$\text{softplus}(a + b) - \text{softplus}(a) = \log \frac{1 + e^{a+b}}{1 + e^a} = \log\big((1 - \sigma(a)) + \sigma(a)\, e^b\big),$$

where $\sigma(a) = e^a/(1 + e^a)$ is the sigmoid. For $a \geq 0$, $\sigma(a) \geq \frac{1}{2}$, and for $b \geq 0$, $e^b \geq 1$, so

$$(1 - \sigma(a)) + \sigma(a)\, e^b = 1 + \sigma(a)(e^b - 1) \;\geq\; 1 + \tfrac{1}{2}(e^b - 1) = \frac{1 + e^b}{2}.$$

Therefore $\mathrm{softplus}(a+b) - \mathrm{softplus}(a) \geq \log \frac{1+e^b}{2} = \mathrm{softplus}(b) - \log 2$. □

Lemma 2 (Range of Δ_j)**.** *For each hidden unit j and all $(\mathbf{v}, \mathbf{c})$, $|\Delta_j(\mathbf{v}, \mathbf{c})| \leq \|\mathbf{U}_j\| \, \|\mathbf{c}\|$. Hence, under Assumption 3, $\mathbb{E}[\Delta_j^2] \leq M_j^2$ with $M_j = \|\mathbf{U}_j\| \, M_c$, and $\mathrm{Var}[\mathrm{CB}] \leq (\sum_j M_j)^2$.*

Proof. Since softplus is monotonically increasing with $\mathrm{softplus}' = \sigma \in (0,1)$, the mean-value theorem gives

$$\begin{aligned} |\Delta_j| &= \left|\mathrm{softplus}\big(\alpha_j^{\mathrm{feat}} + \alpha_j^{\mathrm{ctx}}\big) - \mathrm{softplus}\big(\alpha_j^{\mathrm{feat}}\big)\right| \\ &\leq |\alpha_j^{\mathrm{ctx}}| = |\hat{\mathbf{c}}_j| \\ &\leq \|\mathbf{U}_j\| \, \|\mathbf{c}\|. \end{aligned}$$

Squaring and taking expectations under Assumption 3 gives $\mathbb{E}[\Delta_j^2] \leq \|\mathbf{U}_j\|^2 M_c^2 = M_j^2$, and Minkowski's inequality gives $\mathbb{E}[\mathrm{CB}^2]^{1/2} \leq \sum_j \mathbb{E}[\Delta_j^2]^{1/2} \leq \sum_j M_j$, so $\mathrm{Var}[\mathrm{CB}] \leq (\sum_j M_j)^2$. The sign of Δ_j equals the sign of α_j^{ctx} (by monotonicity of softplus), so both positive and negative values occur. □

Lemma 3 (Lipschitz continuity of CB)**.** *For any $(\mathbf{v}_1, \mathbf{c}_1)$ and $(\mathbf{v}_2, \mathbf{c}_2)$ with $\|\mathbf{c}_1\|_2, \|\mathbf{c}_2\|_2 \leq M_c$,*

$$\left|\mathrm{CB}(\mathbf{v}_1, \mathbf{c}_1) - \mathrm{CB}(\mathbf{v}_2, \mathbf{c}_2)\right| \;\leq\; L_v \, \|\mathbf{v}_1 - \mathbf{v}_2\| + L_c \, \|\mathbf{c}_1 - \mathbf{c}_2\|,$$

where $L_v = \sum_j \|\mathbf{w}_j\|$ and $L_c = \sum_j \|\mathbf{U}_j\|\big(1 + \frac{1}{4}\|\mathrm{Gate}_j\| \, M_c\big)$. In particular, CB is L-Lipschitz with $L = L_v + L_c$ on $\{(\mathbf{v}, \mathbf{c}) \in \mathbb{R}^{2n_v} : \|\mathbf{c}\|_2 \leq M_c\}$.

Proof. For any j, $\nabla_{\mathbf{v}} \Delta_j = \mathbf{w}_j \big[\sigma(\alpha_j^{\mathrm{feat}} + \alpha_j^{\mathrm{ctx}}) - \sigma(\alpha_j^{\mathrm{feat}})\big]$. Since each sigmoid value lies in $(0,1)$, the difference in brackets lies in $(-1,1)$, so $\|\nabla_{\mathbf{v}} \Delta_j\| \leq \|\mathbf{w}_j\|$. By the triangle inequality, $\|\nabla_{\mathbf{v}} \mathrm{CB}\| = \|\sum_j \nabla_{\mathbf{v}} \Delta_j\| \leq \sum_j \|\mathbf{w}_j\| = L_v$.

For the context channel, the GLU gate gives $\alpha_j^{\mathrm{ctx}} = (\mathbf{U}_j^\top \mathbf{c}) \, \sigma(\mathrm{Gate}_j^\top \mathbf{c})$, so

$$\nabla_{\mathbf{c}} \alpha_j^{\mathrm{ctx}} = \mathbf{U}_j \, \sigma(\mathrm{Gate}_j^\top \mathbf{c}) + (\mathbf{U}_j^\top \mathbf{c}) \, \sigma'(\mathrm{Gate}_j^\top \mathbf{c}) \, \mathrm{Gate}_j.$$

Since $\sigma \leq 1$, $\sigma' \leq \frac{1}{4}$, and $|\mathbf{U}_j^\top \mathbf{c}| \leq \|\mathbf{U}_j\| \, M_c$ for $\|\mathbf{c}\|_2 \leq M_c$, $\|\nabla_{\mathbf{c}} \alpha_j^{\mathrm{ctx}}\| \leq \|\mathbf{U}_j\| + \frac{1}{4}\|\mathbf{U}_j\| \, M_c \, \|\mathrm{Gate}_j\|$. Multiplying by $\sigma(\alpha_j^{\mathrm{feat}} + \alpha_j^{\mathrm{ctx}}) \leq 1$ (from $\mathrm{softplus}' = \sigma$) and summing: $\|\nabla_{\mathbf{c}} \mathrm{CB}\| \leq L_c$. Combining via the mean-value theorem on each channel yields the result. □

B.2 Proof of Theorem 1 (Separation)

Proof. The proof proceeds in four steps: decomposition, legitimate lower bound, adversarial upper bound, and tail bounds.

Step 1: Decomposition. Write $\mathrm{CB}(\mathbf{v}, \mathbf{c}) = \sum_{j=1}^{n_h} \Delta_j$ as in (11). We will analyze the per-unit contributions Δ_j separately for the coherent units and their complement.

Step 2: Lower bound for legitimate nodes ($\Psi > 0$). Let $\mathcal{C} \subset \{1, \ldots, n_h\}$ denote the set of *coherent* hidden units (those satisfying Assumption 2), $|\mathcal{C}| \geq \beta_0 n_h$, and for each $j \in \mathcal{C}$ let $E_j = \{\alpha_j^{\mathrm{feat}} \geq 0 \;\wedge\; \alpha_j^{\mathrm{ctx}} \geq 0\}$ be its joint-positive event. On E_j, Lemma 1 (with $a = \alpha_j^{\mathrm{feat}} \geq 0$, $b = \alpha_j^{\mathrm{ctx}} \geq 0$) gives

$$\Delta_j \;\geq\; \mathrm{softplus}(\alpha_j^{\mathrm{ctx}}) - \log 2 \;\geq\; \mathrm{softplus}(0) - \log 2 \;=\; 0. \tag{12}$$

Write $\Delta_j = \Delta_j^+ - \Delta_j^-$ with $\Delta_j^\pm = \max(\pm\Delta_j, 0)$, and define the conditional gains $g_j = \mathbb{E}[\mathrm{softplus}(\alpha_j^{\mathrm{ctx}}) - \log 2 \mid E_j]$ for $j \in \mathcal{C}$, the normalized expected gain $G_{\mathcal{C}} = \frac{1}{n_h} \sum_{j \in \mathcal{C}} \mathbb{P}[E_j] \, g_j$ and the normalized expected negative contribution $N = \frac{1}{n_h} \sum_{j=1}^{n_h} \mathbb{E}[\Delta_j^- \mid \mathcal{P}_{\mathrm{legit}}]$. By (12), $\Delta_j^+ \geq \Delta_j \, \mathbf{1}_{E_j} \geq (\mathrm{softplus}(\alpha_j^{\mathrm{ctx}}) - \log 2) \, \mathbf{1}_{E_j}$ for $j \in \mathcal{C}$, and $\Delta_j^+ \geq 0$ for every unit, so

$$\Psi \;=\; \sum_{j=1}^{n_h} \mathbb{E}[\Delta_j^+ - \Delta_j^- \mid \mathcal{P}_{\mathrm{legit}}] \;\geq\; \sum_{j \in \mathcal{C}} \mathbb{P}[E_j] \, g_j \;-\; \sum_{j=1}^{n_h} \mathbb{E}[\Delta_j^- \mid \mathcal{P}_{\mathrm{legit}}] \;=\; n_h\big(G_{\mathcal{C}} - N\big). \tag{13}$$

Assumption 5 gives $G_{\mathcal{C}} - N \geq \gamma > 0$ with γ independent of n_h, hence

$$\Psi \;=\; \mathbb{E}[\mathrm{CB} \mid \mathcal{P}_{\mathrm{legit}}] \;\geq\; n_h\,\gamma \;=\; \Omega(n_h) \;>\; 0. \tag{14}$$

Step 3: Upper bound for adversarial nodes. For every unit and every input, $\Delta_j \leq \max(\hat{\mathbf{c}}_j, 0)$. If $\hat{\mathbf{c}}_j \leq 0$, monotonicity of softplus gives $\Delta_j \leq 0$. If $\hat{\mathbf{c}}_j > 0$, the mean-value theorem with $\mathrm{softplus}' = \sigma < 1$ gives $\Delta_j \leq \hat{\mathbf{c}}_j$. The bound holds whatever the features, so Assumption 4 gives

$$\Phi \;\leq\; \sum_{j=1}^{n_h} \mathbb{E}_{\mathbf{c}'\sim\mathcal{Q}_{\mathrm{rand}}}\big[\max(\hat{\mathbf{c}}'_j, 0)\big] \;\leq\; \rho_1 M_c \sum_{j=1}^{n_h} \|\mathbf{U}_j\| \;=\; n_h\,\rho_1\,\bar{u}\,M_c. \tag{15}$$

With (14) and the condition $\rho_1 \bar{u} M_c \leq \gamma - \varepsilon$ of Theorem 1,

$$\Psi - \Phi \;\geq\; n_h\big(\gamma - \rho_1 \bar{u} M_c\big) \;\geq\; n_h\,\varepsilon \;=\; \Omega(n_h).$$

The certified margin $n_h(\gamma - \rho_1 \bar{u} M_c)$ decreases linearly as ρ_1 grows.

Degenerate context paths. A unit with $\mathbf{U}_j = \mathbf{0}$ has $\hat{\mathbf{c}}_j \equiv 0$, hence $\Delta_j \equiv 0$, and contributes nothing to $G_{\mathcal{C}}$, N or (15). Units with small $\|\mathbf{U}_j\|$ contribute proportionally little to the adversarial bound.

Feature channel. Feature deviation enters only as a bounded perturbation. At fixed context, the gradient bound $\|\nabla_{\mathbf{v}} \Delta_j\| \leq \|\mathbf{w}_j\|$ from the proof of Lemma 3 gives $|\Delta_j(\mathbf{v}', \mathbf{c}) - \Delta_j(\mathbf{v}, \mathbf{c})| \leq \|\mathbf{w}_j\|_2\, \|\mathbf{v}' - \mathbf{v}\|$, so replacing legitimate features by adversarial features at distance d moves each unit by at most $\|\mathbf{w}_j\|_2\, d$ and the score by at most $L_v\, d$. The bound (15) does not rely on this channel.

Step 4: Tail bounds. Steps 2–3 establish $\Psi - \Phi > 0$. For $\tau^* \in (\Phi, \Psi)$, Chebyshev's inequality gives

$$\mathbb{P}[\mathrm{CB} < \tau^* \mid \mathcal{P}_{\mathrm{legit}}] \;\leq\; \frac{\mathrm{Var}[\mathrm{CB} \mid \mathcal{P}_{\mathrm{legit}}]}{(\Psi - \tau^*)^2}. \tag{16}$$

By Lemma 2, $\mathrm{Var}[\mathrm{CB} \mid \mathcal{P}_{\mathrm{legit}}] \leq (\sum_j M_j)^2 < \infty$. Setting $\tau^* = (\Psi + \Phi)/2$:

$$\mathrm{FPR} \;\leq\; \frac{4\,\mathrm{Var}[\mathrm{CB} \mid \mathcal{P}_{\mathrm{legit}}]}{(\Psi - \Phi)^2}. \tag{17}$$

The adversarial tail follows identically: $1 - \mathrm{TPR} \leq 4\,\mathrm{Var}[\mathrm{CB} \mid \mathcal{P}_{\mathrm{fraud}}, \mathcal{Q}_{\mathrm{rand}}] \,/\, (\Psi - \Phi)^2$. Since $\Psi - \Phi \geq n_h \varepsilon$ (Steps 2–3) and the variances are bounded above by $(\sum_j M_j)^2 = O(n_h^2)$, the bounds are finite for any $\tau^* \in (\Phi, \Psi)$ but need not fall below one. They certify finiteness and do not predict the observed error rates. □

Remark B.1 (Rate under uncorrelated units)**.** If the per-unit contributions $\{\Delta_j\}$ are pairwise uncorrelated under $\mathcal{P}_{\mathrm{legit}}$ (which holds approximately when the weight vectors span diverse directions), then $\mathrm{Var}[\mathrm{CB}] \leq \sum_j \mathrm{Var}[\Delta_j] \leq n_h \overline{M}^2$ with $\overline{M} = \max_j M_j$, and the right-hand side of (17) becomes $O(1/n_h)$. This refinement requires assumptions beyond Assumptions 1–5, and no bound in this paper relies on it.

B.3 Proof of Corollary 1 (Detection Guarantee)

Proof. Set $\tau^* = (\Psi + \Phi)/2$, so $\Psi - \tau^* = \tau^* - \Phi = (\Psi - \Phi)/2 > 0$. Applying (16):

$$\mathrm{FPR} \;\leq\; \frac{4\,\mathrm{Var}[\mathrm{CB} \mid \mathcal{P}_{\mathrm{legit}}]}{(\Psi - \Phi)^2},$$

$$1 - \mathrm{TPR} \;\leq\; \frac{4\,\mathrm{Var}[\mathrm{CB} \mid \mathcal{P}_{\mathrm{fraud}}, \mathcal{Q}_{\mathrm{rand}}]}{(\Psi - \Phi)^2}.$$

Both variances are finite by Lemma 2, and $\Psi - \Phi > 0$ by Theorem 1, so both bounds are finite. □

B.4 Proof of Theorem 2 (Drift Stability)

Proof. **Step 1: Per-step Wasserstein–Lipschitz bound.** By Lemma 3, CB is L-Lipschitz with $L = L_v + L_c$ on the region $\{\|\mathbf{c}\|_2 \le M_c\}$ that carries every $\mathcal{P}_t$ (condition (ii)). Kantorovich–Rubinstein duality gives

$$\left|\mathbb{E}_{\mathcal{P}_t}[\mathrm{CB}] - \mathbb{E}_{\mathcal{P}_{t-1}}[\mathrm{CB}]\right| \;\le\; L \cdot W_1(\mathcal{P}_t, \mathcal{P}_{t-1}) \;\le\; L\,\delta_{\mathrm{drift}},$$

where the first inequality is the dual characterization of W_1 for L-Lipschitz observables and the second is condition (ii).

Step 2: Telescoping. By the triangle inequality,

$$\left|\mathbb{E}[\mathrm{CB}_t] - \mathbb{E}[\mathrm{CB}_0]\right| \;\le\; \sum_{s=1}^{t} \left|\mathbb{E}[\mathrm{CB}_s] - \mathbb{E}[\mathrm{CB}_{s-1}]\right| \;\le\; L\,t\,\delta_{\mathrm{drift}},$$

which establishes (10).

Step 3: Drift horizon for the legitimate mean. Combining Step 2 with the legitimate baseline $\mathbb{E}[\mathrm{CB} \mid \mathcal{P}_0] = \Psi$ gives

$$\mathbb{E}[\mathrm{CB} \mid \mathcal{P}_t] \;\ge\; \Psi - L\,t\,\delta_{\mathrm{drift}}.$$

Condition (i) preserves the assumed feature-context correlation, and $\Psi = \mathbb{E}_{\mathcal{P}_0}[\mathrm{CB}]$ is the initial legitimate baseline. With $\tau^* = (\Psi + \Phi)/2$, the threshold remains strictly below $\mathbb{E}[\mathrm{CB} \mid \mathcal{P}_t]$ whenever

$$L\,t\,\delta_{\mathrm{drift}} \;<\; \tfrac{1}{2}(\Psi - \Phi) \quad\Longleftrightarrow\quad t \;<\; \frac{\Psi - \Phi}{2L\,\delta_{\mathrm{drift}}},$$

the drift horizon (18). Within this horizon, the variance argument of Corollary 1 applies at time t with gap $\mathbb{E}[\mathrm{CB} \mid \mathcal{P}_t] - \tau^*$ (variances remain bounded by Lemma 2). □

Corollary 3 (Drift Horizon)**.** *Under the conditions of Theorem 2, with $\tau^* = (\Psi + \Phi)/2$ as in Corollary 1, the legitimate mean stays above the threshold, $\mathbb{E}[\mathrm{CB} \mid \mathcal{P}_t] > \tau^*$, for any t in the worst-case* drift horizon

$$t \;<\; \frac{\Psi - \Phi}{2L\,\delta_{\mathrm{drift}}}. \tag{18}$$

Proof. Step 3 of the proof of Theorem 2 establishes $\mathbb{E}[\mathrm{CB} \mid \mathcal{P}_t] \ge \Psi - Lt\delta_{\mathrm{drift}}$. This exceeds $\tau^* = (\Psi + \Phi)/2$ exactly when $Lt\delta_{\mathrm{drift}} < (\Psi - \Phi)/2$, i.e., $t < (\Psi - \Phi)/(2L\delta_{\mathrm{drift}})$. The variances remain bounded by Lemma 2, so Chebyshev's inequality still applies at time t, with gap $\mathbb{E}[\mathrm{CB} \mid \mathcal{P}_t] - \tau^*$ in place of $(\Psi - \Phi)/2$. □

B.5 Proof of Corollary 2 (Quantile Stability)

Proof. Let μ_t be the law of the detection score $-\mathrm{CB}(\mathbf{x}_t, \mathbf{c}_t)$ under $\mathcal{P}_t$. Any coupling of $\mathcal{P}_t$ and $\mathcal{P}_0$ maps through $-\mathrm{CB}$ to a coupling of μ_t and μ_0, and $-\mathrm{CB}$ is L-Lipschitz (Lemma 3), so $W_1(\mu_t, \mu_0) \le L\,W_1(\mathcal{P}_t, \mathcal{P}_0)$. The triangle inequality for W_1 and condition (ii) give $W_1(\mathcal{P}_t, \mathcal{P}_0) \le t\,\delta_{\mathrm{drift}}$. On the real line the monotone quantile coupling is optimal, so $W_1(\mu_t, \mu_0) = \int_0^1 |Q_t(u) - Q_0(u)|\,du$, which proves the claim. The bound controls the average displacement of the quantiles, not any single quantile such as the 95th percentile that sets τ^*. □

Remark B.2 (Tightness)**.** The Wasserstein bound is a worst-case upper bound, loose for natural drift, since real drift moves $\mathbf{x}$ and $\mathbf{c}$ jointly along directions where the softplus difference (4) is nearly flat. We discuss this slack quantitatively in §6.5.

Remark B.3 (Coherent Activation justification)**.** Assumption 2 (joint-positive activation for a fraction β_0 of hidden units with probability $\ge p_0$) follows qualitatively from contrastive-divergence training on the correlated $(\mathbf{x}, \mathbf{c})$ pairs of Assumption 1: matched pairs raise both $\alpha_j^{\mathrm{feat}} = \mathbf{w}_j^\top \mathbf{x} + b_j$ and $\alpha_j^{\mathrm{ctx}} = \hat{\mathbf{c}}_j$ on the same training events, since the model is rewarded for assigning low free energy when both channels fire coherently. In the Gaussian-projection regime, the arcsine law [52] relates Assumption 1's correlation ρ_0 to a sign-correlation lower bound $\frac{2}{\pi}\arcsin(\rho_0) \ge \frac{2}{\pi}\rho_0$, which under symmetric pre-activation marginals translates to $p_0 \ge \frac{1}{4} + \frac{\rho_0}{2\pi}$. The constant β_0 is measured after model convergence and p_0 is bounded by the arcsine argument above.

Table 2: Verified Assumption 1–3 constants on training data.

Dataset	ρ_0 (A1)	β_0 (A2)	M_x (A3)	M_c (A3)
XBlock	0.30	0.38	6.6	23.5
Reddit	0.68	0.95	8.6	7.8

C Datasets

XBlock [21] is an Ethereum transaction graph with 153,477 wallet nodes (1,127 labeled phishing, 0.7%) and 715,573 directed transaction edges. Node features are 38 per-wallet transaction statistics (counts, incoming/outgoing volumes, gas usage, temporal spread). The label-graph correlation is anti-homophilic. The sign-corrected feature-context correlation is $\rho_0 = 0.30$ (Eq. (1)). **Reddit** [22] is a social-media graph with 10,984 user nodes (366 banned users, 3.3%) and 168,016 edges between users who co-participate in subreddits. Features are 73-dimensional, combining 64 LIWC linguistic categories computed over each user's posts with 9 graph-structural signals (degree, clustering, centrality variants). Labels correlate homophilically with the graph. The sign-corrected feature-context correlation is $\rho_0 = 0.68$. Features are globally normalized. A 10% held-out split of the pool assumed legitimate, which the datasets' benign labels define in our experiments, is used for threshold calibration and drift-induced FPR measurement, and a fixed attack-generation reserve of $\sim$200 labeled adversaries is held out per dataset, so the supervised baselines see the remaining 932 / 166 adversary labels; all other nodes are used for training. Without drift, the Reddit threshold is stable down to small calibration samples, with FPR $5.9 \pm 0.6\%$ at 530 calibration nodes and $6.0 \pm 0.2\%$ at 1,061 over three sampling seeds, and Isolated Injection TPR $93.7 \pm 0.2\%$ in that run.

Verified theorem assumptions. The constants entering Theorems 1–2 are measured on training data: $\rho_0 = \frac{1}{N}\sum_i \cos(\mathbf{x}_i, \mathbf{c}_i)$ on the pool assumed legitimate before training; β_0, M_x, M_c after model convergence (Table 2). M_x and M_c are root-mean-square (RMS) norms over that pool, and the drift budgets of Table 12 evaluate L at this value of M_c rather than at a verified support radius. Assumption 5 holds for the trained models on both datasets.

The asymmetry $M_c > M_x$ on XBlock reflects sign-corrected aggregation of positive-mean transaction features under a heavy-tailed degree distribution; Reddit's centered LIWC features yield a balanced pair.

D Additional Baselines

Protocol. Every detector in Tables 3 and 4 is scored on the same legitimate evaluation nodes and the same attack reserve, with the attacks of §3. Under drift, models and thresholds stay frozen at t=0, with the single exception of ARC described below, and the Combined Score follows §6.1.

Implementations. The nine detectors of §6.1 follow their published configurations, and model selection uses a held-out validation split of the training pool. XGBGraph, BWGNN and Graph Transformer are the GADBench models [46], and GAD-NR, CONAD, AnomalyDAE and OCGNN run through PyGOD [53]. CONAD and AnomalyDAE exhausted memory on XBlock's 153K nodes under this configuration and are reported on Reddit only. GDN and ARC run from their official repositories. ARC is a pretrained generalist detector applied without fine-tuning, with 300 legitimate nodes as its in-context normal set. It re-runs its per-graph preprocessing on every evaluated graph, which is its intended usage, while its threshold stays frozen at t=0. Fixing that preprocessing at t=0 as well gives Combined Scores of 0.27 on XBlock and 0.46 on Reddit.

Ablations. The signed-cosine score is $-\cos(\mathbf{x}_i, \mathbf{c}_i)$ with the sign-corrected context, the feature-context principle with no learning. Context-MLP is a two-layer MLP trained on legitimate pairs to predict $\mathbf{x}$ from $\mathbf{c}$ and scored by $\|\mathbf{x} - \hat{\mathbf{x}}(\mathbf{c})\|_2$, which keeps context conditioning without the free-energy differential or the GLU gate.

Table 3: All detectors at the last drift step, $T=20$.

		XBlock			Reddit		
Family	Method	TPR	$\text{FPR}_{\text{drift}}$	Combined	TPR	$\text{FPR}_{\text{drift}}$	Combined
Ours	CB-RBM	0.99	5.2	**0.94**	0.96	8.4	**0.88**
Supervised	GraphSAGE	0.89	5.6	0.84	0.06	5.0	0.05
	GAT	0.71	11.6	0.63	0.32	4.6	0.31
	GCN	0.75	5.4	0.71	0.36	4.6	0.35
	XGBGraph	0.98	17.8	0.81	0.01	0.0	0.01
	Graph Transformer	0.87	11.1	0.77	0.07	3.8	0.07
	BWGNN	0.54	5.7	0.51	0.17	8.2	0.16
Shift-robust	GDN	0.77	10.4	0.69	0.24	1.8	0.23
Generalist	ARC	0.40	7.0	0.37	0.86	47.2	0.45
Reconstruction	DOMINANT	0.84	99.0	0.01	0.81	43.2	0.46
	GAE	0.84	66.6	0.28	0.80	91.8	0.07
	GAD-NR	0.53	3.4	0.51	0.67	5.4	0.63
	AnomalyDAE	–	–	–	0.60	82.0	0.11
Contrastive	CoLA	0.19	15.4	0.16	0.56	8.6	0.51
	CONAD	–	–	–	0.79	36.4	0.50
One-class	OCGNN	0.20	7.0	0.18	0.46	7.0	0.43
Ablation	Std-RBM (no context)	0.41	99.8	0.001	0.09	89.4	0.010
	Signed cosine (no learning)	0.18	7.7	0.17	0.86	33.4	0.57
	Context-MLP (no RBM/GLU)	0.76	29.9	0.53	0.92	95.7	0.04

Table 4: Per-attack AUROC / TPR@5%FPR (%).

	Method	Isolated Inj.	Feat. Camo	Sybil	Rel. Camo
XBlock	XGBGraph	99.9 / 99.3	98.9 / 93.2	99.9 / 99.7	99.7 / 99.9
	Graph Transformer	98.5 / 94.2	92.7 / 71.6	98.7 / 95.2	97.1 / 86.2
	BWGNN	83.4 / 28.5	65.1 / 3.0	99.1 / 97.1	97.8 / 88.0
	GDN	95.8 / 90.3	90.2 / 79.0	99.1 / 96.2	65.4 / 44.3
	ARC	94.8 / 40.0	94.9 / 40.0	92.3 / 40.0	91.0 / 39.5
	GAD-NR	99.7 / 100.0	99.4 / 100.0	60.1 / 0.0	82.0 / 12.3
	OCGNN	77.3 / 34.9	68.9 / 38.0	55.5 / 5.1	61.9 / 0.5
	Signed cosine	83.0 / 19.5	84.1 / 23.1	67.2 / 15.9	70.9 / 13.3
	Context-MLP	97.1 / 81.5	96.1 / 77.4	97.6 / 90.3	93.4 / 54.9
	AnomalyDAE	–	–	–	–
	CONAD	–	–	–	–
Reddit	XGBGraph	59.3 / 0.0	59.3 / 0.0	61.8 / 0.0	64.1 / 2.5
	Graph Transformer	67.0 / 0.7	60.7 / 0.0	66.7 / 0.2	82.8 / 28.3
	BWGNN	62.0 / 7.7	58.0 / 7.9	57.2 / 11.0	87.5 / 42.3
	GDN	67.5 / 18.8	56.1 / 9.5	88.3 / 53.3	67.4 / 14.0
	ARC	100.0 / 100.0	100.0 / 100.0	97.9 / 92.1	89.7 / 50.1
	GAD-NR	100.0 / 100.0	100.0 / 100.0	91.0 / 55.5	77.3 / 11.5
	AnomalyDAE	95.2 / 88.5	97.0 / 90.5	81.8 / 46.5	65.0 / 15.5
	CONAD	97.7 / 93.0	88.2 / 77.5	99.8 / 100.0	92.0 / 43.5
	OCGNN	94.5 / 75.5	96.3 / 87.0	67.6 / 0.0	78.3 / 23.0
	Signed cosine	100.0 / 100.0	100.0 / 100.0	90.5 / 71.5	88.9 / 71.5
	Context-MLP	99.5 / 100.0	99.3 / 100.0	98.1 / 97.0	95.4 / 69.0

E Robustness Experiments

E.1 Contaminated training pool

We mix fraud into the pool assumed legitimate and refit the sign vector, the model and the threshold on that pool, so the whole pipeline stays label-free. The legitimate nodes are split into training (80%), calibration (10%) and held-out (10%) sets. The natural level adds every non-reserve fraud node once, which is 0.8% of the training pool on XBlock and 1.9% on Reddit, and contaminates the calibration set at the same rate. The 5% and 10% levels oversample fraud to that share of both sets. The clean row is this protocol's own reference, refit on the 80% split, and is not the Table 1 model. FPR is measured on the held-out legitimate nodes at the threshold calibrated on the contaminated set, and the attacks follow §3. No sign dimension flips at any level on either dataset, and Isolated Injection, Feature Camouflage and Sybil stay at 100% TPR on XBlock and at least 92% on Reddit.

Table 5: Fraud in the training and calibration pools (FPR and TPR in %).

	XBlock			Reddit		
Fraud in pool	FPR	AUROC	Rel. Camo TPR	FPR	AUROC	Rel. Camo TPR
0% (clean)	4.6	0.99	80.5	4.0	0.98	65.5
Natural	4.3	0.99	85.6	4.5	0.98	63.5
5%	3.4	0.99	68.2	4.6	0.98	70.0
10%	2.7	0.99	39.0	5.0	0.98	66.0

Table 6: Sign vector with a share of its dimensions flipped, and after re-estimating it without labels (FPR and TPR in %).

	XBlock			Reddit		
Flipped	FPR	Iso. TPR	FPR after refresh	FPR	Iso. TPR	FPR after refresh
0%	5.0	100.0	5.0	5.0	97.5	5.0
10%	19.1	100.0	5.0	11.6	99.0	5.0
25%	94.4	100.0	5.0	93.4	98.0	5.0
50%	99.9	100.0	5.0	97.5	97.5	5.0

E.2 Sign vector

We test a stale sign vector by flipping a random 10%, 25% or 50% of its dimensions while the model and the threshold stay fixed at their clean values, and then re-estimate $\mathbf{s}$ from feature statistics on the current unlabeled nodes, exactly as it was first fitted (Table 6). The refresh recovers all 38 XBlock and all 73 Reddit dimensions and restores Isolated Injection TPR to 100% and 97.5%. We also estimate $\mathbf{s}$ from random samples of unlabeled legitimate nodes, with the model and the threshold again fixed (Table 7). Isolated Injection TPR stays at 100% on XBlock and at least 97.5% on Reddit at every sample size. On Reddit, samples of 53 or 106 nodes still get some signs wrong in some seeds, so FPR varies widely across seeds, while from 1,061 nodes every seed recovers all signs.

F Feature-Context Decomposition

To identify which signal CB-RBM relies on, we score each test node under four controlled conditions that independently swap features and context between legitimate and adversarial/random sources. Condition A pairs adversarial features with random context; Condition B pairs legitimate features with the same random context; Condition C uses legitimate features with legitimate context (the null condition); Condition D pairs adversarial features with legitimate context. Results are summarized in Table 8.

Two regularities appear on both datasets. First, A≈B: random context triggers detection regardless of whether the features are adversarial or legitimate, confirming that CB primarily flags absent context rather than feature deviation. Second, Condition C is near zero, so legitimate features paired with legitimate context evade detection. The datasets diverge on Condition D. On XBlock, D≈C: legitimate context alone suppresses the score, and the feature channel contributes nothing measurable. On Reddit, D is substantially above C, so adversarial features with legitimate context are partially flagged. This residual feature sensitivity is consistent with Reddit's richer feature space (73 LIWC and structural features versus XBlock's 38 transaction statistics): the learned weights carry more discriminative feature signal, which surfaces when the context channel is inactive. The main-body interpretation (Section 6.5) still holds — CB-RBM is fundamentally a network-context-presence detector, with feature sensitivity acting only as a secondary channel when the feature space is informative on its own.

Unit-level trace. Table 9 follows single nodes through the trained models. A coherent unit has positive feature and gated context pre-activations, so it adds a positive increment to CB, and the last column averages the gated context pre-activation over these units. The legitimate node retains many coherent units and its gated context contribution after 20 drift steps, so its CB stays high. The injected

Table 7: Sign vector estimated from n unlabeled legitimate nodes (FPR in %). $\pm$ is the standard deviation over 5 seeds.

XBlock (38 dimensions)			Reddit (73 dimensions)		
n	Wrong dimensions	FPR	n	Wrong dimensions	FPR
763	1.4 ± 0.8	6.9 ± 1.7	53	1.0 ± 0.9	28.7 ± 33.8
1,526	0.8 ± 0.8	6.5 ± 2.8	106	1.0 ± 0.9	28.7 ± 33.8
7,630	0.6 ± 0.5	5.1 ± 0.1	530	0.2 ± 0.4	6.0 ± 2.0
15,260	0.4 ± 0.5	5.1 ± 0.1	1,061	0.0 ± 0.0	5.0 ± 0.0
152,600	0.0 ± 0.0	5.0 ± 0.0	10,618	0.0 ± 0.0	5.0 ± 0.0

Table 8: Feature-context decomposition, Combined Score.

Condition	Features	Context	XBlock	Reddit
A	adversarial	random	0.942	0.921
B	legitimate	random	0.942	0.929
C	legitimate	legitimate	0.078	0.095
D	adversarial	legitimate	0.111	0.632

adversary has random context, the gate shrinks its context contribution, and its CB drops to 28.4 on XBlock and 33.8 on Reddit.

G Ablation

This appendix examines three design choices: the context module itself, the hidden-layer capacity n_h, and the neighborhood-pooling operator used to construct the raw context vector $\mathbf{c}$.

Context module. Removing the context pathway reduces CB-RBM to Std-RBM, a standard RBM scored by free energy alone. Combining the attack-benchmark TPR (Table 1) with the T=20 drift FPR of Table 3 yields a mean Combined Score of ≈ 0.001 on XBlock and ≈ 0.010 on Reddit for Std-RBM, compared with 0.94 and 0.88 for CB-RBM. The collapse spans two to three orders of magnitude on both datasets and localizes the entire empirical benefit to the context channel.

Hidden-layer capacity. Figure 5 shades the min–max envelope of Combined Score across all four attacks as n_h varies from 32 to 1024. A tight band at high Combined Score indicates a capacity where every attack is simultaneously well-detected; a wide or low band indicates a regime where at least one attack slips through. The lower boundary of the envelope is essentially Relation Camouflage — the attack with the largest ρ_1^{eff} and the smallest measured separation gap (Figure 2). The envelope rises with n_h until the effective coherent-unit count $\beta_0 n_h$ (Assumption 2, measured on training data) saturates, and beyond saturation further capacity only over-fits the legitimate distribution. The two datasets diverge at this saturation point: XBlock's envelope keeps rising through n_h=256 as the larger pool ($\sim$152K legitimate nodes) absorbs additional capacity, while Reddit's envelope peaks at n_h=128 and degrades sharply for $n_h \geq 256$ as Relation Camouflage Combined Score collapses from 0.81 to 0.33. The main body accordingly uses n_h=256 for XBlock and n_h=128 for Reddit.

Neighborhood pooling. The raw context vector $\mathbf{c} = \text{Pool}\{\mathbf{x}_u : u \in \mathcal{N}(v)\}$ is constructed by pooling neighbor features. Table 10 compares four operators on XBlock under two attacks chosen to bracket the attack-difficulty spectrum: Isolated Injection (random, unstructured context — the easiest case for CB-RBM) and PGD (white-box gradient attack perturbing $\mathbf{v}$ and $\mathbf{c}$ jointly — the hardest). We report only this pair because the remaining static attacks in Table 1 (Feature Camouflage, Sybil, Relation Camouflage) also rely on unstructured, Sybil-generated or hub-borrowed context and track Isolated Injection closely, adding no new information about the pooling operator; the contrast with PGD is what stress-tests the choice. We run this ablation only on XBlock because XBlock uses raw transaction statistics as features, so the pooling operator is the sole mechanism by which neighborhood information enters the model. Reddit features already include engineered graph-derived structural signals (9 of its 73 dimensions), so some neighborhood information is available to the model

Table 9: Unit-level trace through the trained models.

Dataset	Node	CB	Coherent units	Mean gated context
XBlock	Legitimate (median CB)	1046.0	111/256	+7.36
	Same node after 20 drift steps	952.0	88/256	+6.51
	Injected adversary	28.4	97/256	+0.55
Reddit	Legitimate (median CB)	116.2	57/128	+2.41
	Same node after 20 drift steps	115.5	52/128	+2.45
	Injected adversary	33.8	34/128	+1.39

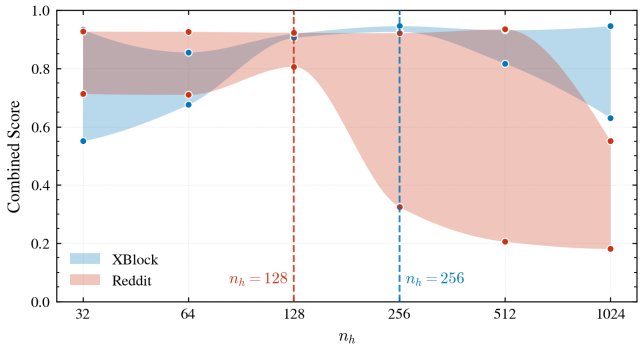


Figure 5: Capacity ablation. Shaded band shows min–max Combined Score across all four attacks at each n_h. Dashed vertical lines mark the optimal values: n_h=256 for XBlock, n_h=128 for Reddit.

independently of how $\mathbf{c}$ is constructed, making the pooling operator less decisive on that dataset. Mean and attention pooling tie at Combined $= 0.95$ on Isolated Injection; max and weighted-degree trail slightly. Under PGD, mean and attention retain a meaningful score while max collapses to 0.15. All four operators preserve per-dimension alignment (they return a vector in the original feature space rather than a learned embedding), which is what the sign vector $\text{sign}(\text{Pearson}(\mathbf{x}_d, \mathbf{c}_d))$ requires to identify the legitimate correlation pattern.

H PGD Budget Sweep

This appendix evaluates CB-RBM under a white-box threat model: the adversary has gradient access to the detector's weights $\mathbf{W}$, $\mathbf{U}$, and the gate, and jointly optimizes $\mathbf{v}$ and $\mathbf{c}$ under an ℓ_∞-budget. Starting from real adversarial nodes at their natural graph positions, we sweep the perturbation budget $\epsilon \in \{0.01, 0.05, 0.1, 0.2, 0.3, 0.5\}$ with 50 projected-gradient steps. Figure 6 plots detection AUROC on both datasets.

Detection degrades monotonically with ϵ on both datasets. On XBlock, AUROC falls from 0.90 at ϵ=0.01 to 0.66 at ϵ=0.5, staying above random-chance across the full range. On Reddit, AUROC drops to 0.42 already at ϵ=0.1 and continues to 0.07 at ϵ=0.5; the starting AUROC is lower here because in-graph adversaries reach CB-RBM through the feature channel rather than the context-absence channel. Two factors explain the faster Reddit collapse. First, Reddit's 73-dimensional feature space gives the PGD optimizer more independent directions in which to move $\mathbf{v}$ under a fixed ℓ_∞ budget than XBlock's 38-dimensional space. Second, Reddit's higher base correlation $\rho_0 = 0.68$ means the learned GLU gate is tuned to strong context–feature agreement, so small coordinated perturbations to $\mathbf{v}$ and $\mathbf{c}$ move the CB score further per unit of perturbation than on XBlock. These results are consistent with the Limitations (§6.5): CB-RBM is designed as a low-cost first-line detector for the common operational case of an adversary without white-box gradient access; combining it with an adversarially trained or certified-robustness second stage would be the natural defense under the PGD threat model.

Table 10: Pooling operator on XBlock, Combined Score.

Attack	Mean	Max	Weighted-deg.	Attention
Isolated Injection	0.950	0.914	0.930	0.950
PGD	0.326	0.145	0.224	0.263

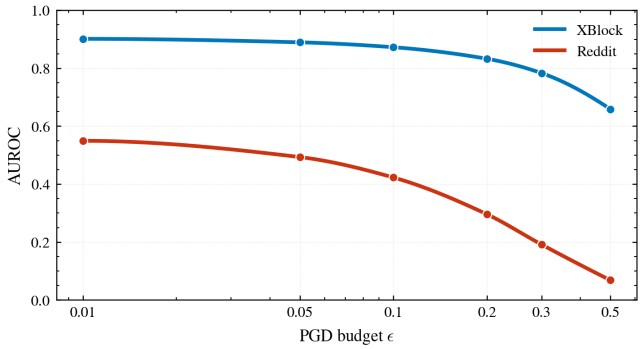


Figure 6: White-box PGD budget sweep. Detection AUROC vs. perturbation budget ϵ.

I Drift Stability Extensions

The main body reports Combined Score under a random-walk drift simulation, following standard practice in the concept-drift literature [8, 28]. This appendix extends that analysis with alternative drift regimes that vary tails and correlation structure, threshold-level and correlation-weakening drift, real temporal evolution on the XBlock 2017–2018 transaction data, and adversarial-trajectory detection latency (§I.1–§I.5).

I.1 Alternative drift regimes

At each step t, a drift increment $\boldsymbol{\eta}_t$ is added to every node's features and the context vector is recomputed on the updated graph, so drift propagates jointly through node attributes and their aggregated neighborhood — matching how real distributional shift unfolds on evolving graphs such as the wallet-turnover and user-behavior patterns of the main-body domains. We compare five increment distributions chosen to cover the natural-drift regimes a detector is likely to encounter in deployment: *Gaussian* — isotropic bounded-variance drift in the same family as the main-body simulation; *correlated* ρ=0.3 and *correlated* ρ=0.6 — coordinated population shifts where subsets of features co-move, modeled with equicorrelated Gaussian increments; and *heavy-tail* df=10 / *heavy-tail* df=3 — bursty, outlier-driven drift modeled with Student-t increments, smaller df yielding heavier tails. Table 11 reports the drift-induced false-positive rate at T=20 for CB-RBM and Std-RBM on both datasets. The per-step variance budget here is calibrated for cross-regime comparison and differs from the headline drift simulation; the qualitative pattern — Std-RBM collapse, CB-RBM stability — is unchanged.

CB-RBM stays at $\leq 16\%$ FPR across every regime and dataset, while Std-RBM exceeds 55% in every regime and $> 95\%$ in four of the five XBlock regimes. Two observations are consistent with Theorem 2. First, the Wasserstein–Lipschitz bound is *distribution-agnostic*: it depends only on $W_1(\mathcal{P}_t, \mathcal{P}_{t-1}) \leq \delta_{\mathrm{drift}}$, not on the tails or correlation structure of the drift increment, consistent with CB-RBM behaving similarly across all five variants (the slightly lower df=3 row reflects Student-t's variance-matched mass concentrating near zero, so most nodes drift less than under Gaussian rather than more). Second, the two datasets respond oppositely to correlated drift by the same mechanism. Drift aligned with the learned $(\mathbf{v}, \mathbf{c})$ correlation lies in a near-flat valley of the CB surface — the softplus differences in Eq. (4) nearly cancel — so FPR drops; drift off this axis avoids the valley and FPR rises. Reddit's strong ρ_0=0.68 aligns closely with the equicorrelated ρ=0.6 increment, driving FPR toward zero ($9.0 \to 0.2\%$); XBlock's weaker ρ_0=0.30 lies further off-axis, raising FPR modestly ($8.4 \to 15.4\%$, still far below Std-RBM at every regime). This is consistent with the "conservative bound" discussion of Section 6.5: real drift's CB-impact depends on alignment with the learned correlation, while the worst-case Theorem 2 bound holds in either direction.

Table 11: Drift-induced FPR at $T=20$ under five drift regimes (%).

	XBlock		Reddit	
Regime	CB-RBM	Std-RBM	CB-RBM	Std-RBM
Gaussian	8.4	96.4	9.0	73.8
Correlated $\rho=0.3$	12.8	95.8	2.8	74.0
Correlated $\rho=0.6$	15.4	97.2	0.2	69.0
Heavy-tail df=10	10.8	95.2	9.0	71.2
Heavy-tail df=3	7.6	73.8	6.6	55.2

Table 12: Wasserstein–Lipschitz bound vs. observed CB shift on legitimate nodes (Gaussian drift). L is the Lipschitz constant of CB evaluated at the RMS value of M_c (Appendix C); δ_{drift} is the per-step W_1 budget; the RMS-based budget is $L \cdot T \cdot \delta_{\text{drift}}$; slack is budget/observed.

Dataset	L	δ_{drift}	T	RMS-based budget	Observed $\lvert\Delta\text{CB}\rvert$	Slack
XBlock	1189.5	1.23	5	7,288	23.3	312×
			10	14,575	35.7	408×
			15	21,863	46.1	474×
			20	29,151	55.4	**526**×
Reddit	154.0	1.70	5	1,313	0.68	1,921×
			10	2,625	1.03	2,549×
			15	3,938	1.17	3,373×
			20	5,250	1.16	**4,540**×

RMS-based numerical comparison. Table 12 grounds the slack-factor numbers reported in §6.5. L is the Lipschitz expression of Lemma 3 evaluated at the RMS value of M_c (Appendix C); δ_{drift} is the per-step Wasserstein-1 budget; observed $|\Delta\text{CB}|$ is the realized mean-CB shift on legitimate nodes after T Gaussian-drift steps. The slack factor (bound/observed) grows with T on both datasets and is consistent with the conservative-bound argument: the worst-case $L \cdot T \cdot \delta_{\text{drift}}$ allows every coordinate to move adversarially, while real drift moves $(\mathbf{v}, \mathbf{c})$ jointly along the legitimate-correlation axis where the softplus differences nearly cancel.

I.2 Threshold-level drift

Evaluating $L\,T\,\delta_{\text{drift}}$ with the RMS-based L gives 29,151 on XBlock and 5,250 on Reddit at $T=20$ (Table 12). The threshold is instead set by the 95th percentile of the legitimate scores, which we track in a separate three-seed run with the model and the threshold frozen at $t=0$ (Table 13). At $T=20$ this percentile shifts by 42.7 on XBlock and 3.4 on Reddit, about a thousandth of that budget, and mean FPR at the frozen threshold rises from 5% to 6.1% and 7.1%. Using the same RMS-based value of L, the horizon expression of Corollary 3 evaluates to 0.32 steps on XBlock and 0.12 on Reddit, while mean FPR at the frozen 95th-percentile threshold stays within 2.1 points of the target at every reported step through $T=20$.

I.3 Correlation-weakening drift

The drift of Figure 3 moves features and context together. We weaken that link in three ways at $T=20$, with the model and the threshold frozen at $t=0$. Context lag recomputes each node's context from features L steps behind its own. Private drift adds to each node an independent drift that its neighbors' context does not include, scaled by p relative to the shared drift. Reversal negates each node's drift relative to its neighborhood on a fraction of the dimensions, so features and context anti-align there (Table 14). Recalibrating the threshold at the 95th percentile of the drifted pool's scores, the label-free refresh of Algorithm 1, returns FPR to 5% by construction, and under reversal it keeps Isolated Injection TPR at 100% on XBlock and at least 95.5% on Reddit. Table 15 applies the same recalibration to the eight methods of Table 1 under the co-evolving drift of Figure 3 and scores the four attacks against the recalibrated threshold.

Table 13: Shift of the 95th-percentile threshold and FPR (%) at the threshold frozen at t=0. $\pm$ is the standard deviation over 3 drift seeds.

	XBlock		Reddit	
T	Threshold shift	FPR	Threshold shift	FPR
1	5.6 ± 1.9	5.1 ± 0.1	0.18 ± 0.01	5.1 ± 0.0
5	14.1 ± 4.2	5.4 ± 0.1	0.84 ± 0.23	5.5 ± 0.1
10	27.3 ± 5.8	5.6 ± 0.1	1.67 ± 0.31	6.0 ± 0.1
15	34.2 ± 6.6	5.8 ± 0.1	2.55 ± 0.27	6.4 ± 0.2
20	42.7 ± 5.9	6.1 ± 0.1	3.41 ± 0.27	7.1 ± 0.3

Table 14: Correlation-weakening drift at T=20, measured ρ and legitimate FPR (%) at the frozen threshold. Each block has its own co-evolving reference run.

	XBlock		Reddit	
Regime	ρ	FPR	ρ	FPR
Co-evolving (reference)	0.141	5.4	0.615	7.5
Context lag, L=5	0.141	5.3	0.588	8.2
Context lag, L=10	0.142	5.3	0.567	9.4
Context lag, L=20	0.144	5.6	0.509	10.6
Private drift, p=0.3	0.138	5.4	0.602	8.0
Private drift, p=0.6	0.125	5.7	0.568	9.4
Private drift, p=1.0	0.113	6.2	0.506	13.2
Co-evolving (reference)	0.146	6.1	0.613	7.8
Reversed, 10% of dims	0.145	6.1	0.581	8.1
Reversed, 25% of dims	0.144	6.2	0.534	8.5
Reversed, 50% of dims	0.141	6.1	0.464	10.1

Topology change. We also change the edges at T=20, on top of the feature drift, with the model and the threshold frozen at t=0. Growth keeps every edge and adds 10%, 30% or 50% new triadic edges, the pattern of cumulative interaction graphs such as ours, where on-chain edges are permanent and Reddit interactions accumulate. It raises the legitimate FPR to 12.6, 24.7 and 37.3% on XBlock and 12.9, 24.6 and 34.6% on Reddit. The label-free threshold refresh restores 5.0% on both, with Isolated Injection TPR at 100% on XBlock and 95.5, 93.0 and 91.0% on Reddit. Random rewiring of 10% of edges, evaluated as a topology stress test, raises the FPR to 20.4% and 31.7%, and the refresh restores 5.0% with Isolated Injection TPR at 76.4% and 89.5%.

I.4 Real temporal validation (XBlock)

To check that synthetic drift is not the only regime where CB-RBM holds up, we train on XBlock data through 2017 (a dataset with natively timestamped edges) and evaluate on five cumulative windows, the 2017 training window and 2018-Q1 through Q4. Between windows the Ethereum ecosystem evolves organically — new wallets appear, transaction patterns shift, and each window adds roughly 200K–500K new edges over the preceding one — producing real natural drift rather than a simulated increment. Figure 7 plots FPR and TPR at the calibrated 5%-FPR operating point for all eight methods across the five windows. Node features in each window use transactions up to its cutoff only, and the scaler, sign vector and threshold are fitted on the training window and then frozen. Rebuilding each window's context from pre-cutoff edges only gives FPR 5.0, 3.8, 3.4, 3.2 and 3.1% across the five windows, against 5.0, 3.5, 3.2, 3.2 and 3.1% for the construction used in Figure 7.

CB-RBM holds at TPR= 100% across all five windows; the runner-up (GAT) holds around 84%–89%. The feature distribution drifts monotonically — $\|\boldsymbol{\mu}_t - \boldsymbol{\mu}_0\|$ grows from 0 to $\approx$ 73 over the windows — without translating into an FPR increase or TPR drop for CB-RBM. The supervised GNN baselines are also stable at lower absolute levels ($\approx$ 70%–89%), while Std-RBM and GAE hold flat in the 45–55% band, DOMINANT trends slightly higher (up to $\approx$ 60%), and CoLA stays below 16%. Real temporal evolution therefore reproduces the pattern of the synthetic drift simulation.

Table 15: Threshold recalibration for the eight methods of Table 1 under the co-evolving drift at T=20, Combined Score before and after, from this experiment's run. Best refreshed score in bold.

	XBlock		Reddit	
Method	Frozen	Refreshed	Frozen	Refreshed
CB-RBM	0.92	**0.90**	0.88	**0.90**
Std-RBM	0.003	0.004	0.016	0.014
GraphSAGE	0.82	0.82	0.06	0.06
GAT	0.60	0.61	0.11	0.11
GCN	0.74	0.72	0.15	0.15
DOMINANT	0.01	0.02	0.53	0.64
CoLA	0.19	0.07	0.16	0.13
GAE	0.32	0.01	0.05	0.48

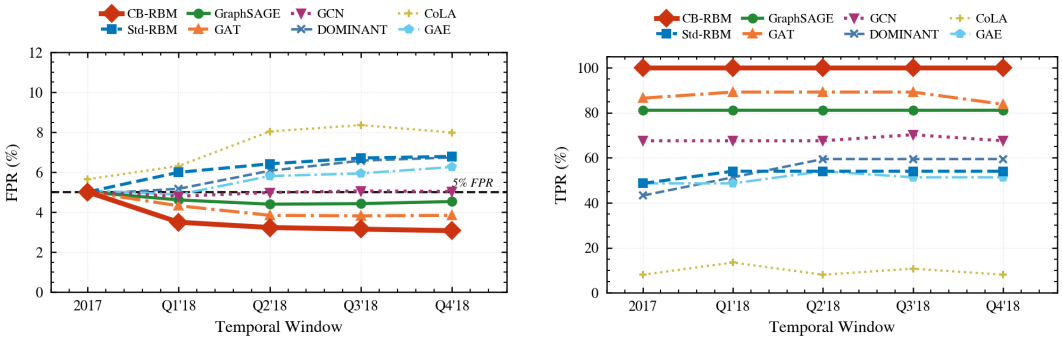


Figure 7: XBlock real-temporal validation. (Left) FPR and (right) TPR at the calibrated 5%-FPR operating point across five windows (trained and calibrated on the 2017 window, evaluated on 2017–2018 windows).

CB-RBM and the supervised GNNs remain near the calibration target, while the reconstruction and density baselines show increased false alarms.

I.5 Detection latency

To complement the aggregate FPR view, Table 16 reports the trajectory step T at which each detector first fires on a five-scenario suite: one benign control (natural drift, which should *not* trigger an alarm) and the four adversarial scenarios of Table 1. The setup evolves 200 target nodes over T=0, . . . , 50 under each scenario; the *first-fire* step is the earliest T at which at least 90% of those nodes exceed the detection threshold. "Never" means the 90% condition is not triggered within T=50 steps.

The pattern is the dual of the drift-stability view. Std-RBM false-alarms on natural drift at T=15 (XBlock) / T=23 (Reddit) and is blind to every adversarial trajectory: the free-energy channel alone cannot distinguish the attacks from legitimate drift once context is excluded. CB-RBM shows the opposite pattern, staying silent under natural drift and firing on every adversarial trajectory. Sybil and Relation Camouflage are fixed-state attacks (the clique injection or hub rewire is present from T=0), and CB-RBM fires immediately. Isolated Injection and Feature Camouflage are evolving trajectories: context interpolates from legitimate at T=0 toward random by T=50, so CB-RBM (correctly) does not fire while the context still resembles legitimate input, and the T=18–26 first-fire step marks the crossing where the context has decayed enough to become distinguishable. Two-channel latency therefore separates cleanly: the feature-only model fires on the wrong trajectory, while the context-aware model fires on the right ones.

Growing Sybil. The Sybil attack of Table 1 uses a clique of fixed size from the start. Here the clique size grows from 2 to 50 over 50 steps. At each step the clique is rebuilt at its current size and the context of the 195 XBlock and 200 Reddit attack nodes connected to it is recomputed, while each detector keeps the threshold calibrated at t=0 (Table 17). With two members CB-RBM already flags 92.3% of the attack nodes on XBlock and 61.0% on Reddit, crosses 90% on Reddit by step 4

Table 16: First-fire step T (first T with $\geq 90\%$ of 200 target nodes flagged). Lower is better for adversarial rows; "never" is correct for natural drift.

	XBlock		Reddit	
Scenario	CB-RBM	Std-RBM	CB-RBM	Std-RBM
Natural drift (benign)	never	15	never	23
Sybil	0	never	0	never
Relation Camouflage	0	never	0	never
Isolated Injection	18	never	24	never
Feature Camouflage	24	never	26	never

Table 17: Fraction of attack nodes flagged (%) as the Sybil clique grows from 2 to 50 members.

		XBlock		Reddit	
Step	Clique size	CB-RBM	Std-RBM	CB-RBM	Std-RBM
0	2	92.3	51.3	61.0	9.5
10	12	100.0	51.3	100.0	9.5
20	21	100.0	51.3	100.0	9.5
30	31	100.0	51.3	100.0	9.5
40	40	100.0	51.3	100.0	9.5
50	50	100.0	51.3	100.0	9.5

(between the rows of the table) and flags 100% on both datasets by step 10. Std-RBM stays at 51.3% and 9.5% throughout.

J Context Interpolation Sweep

We complement Table 1's discrete attacks with a *context interpolation sweep*: a continuous-intensity extension of the camouflage concept introduced in the fraud-detection literature [54]. An attacker parameter $\alpha \in [0, 1]$ interpolates between the more *revealing* and the more *camouflaging* context for the dataset:

$$c_\alpha \;=\; (1-\alpha)\, c_{\text{revealed}} + \alpha\, c_{\text{camo}}, \qquad \alpha \in [0,1], \tag{19}$$

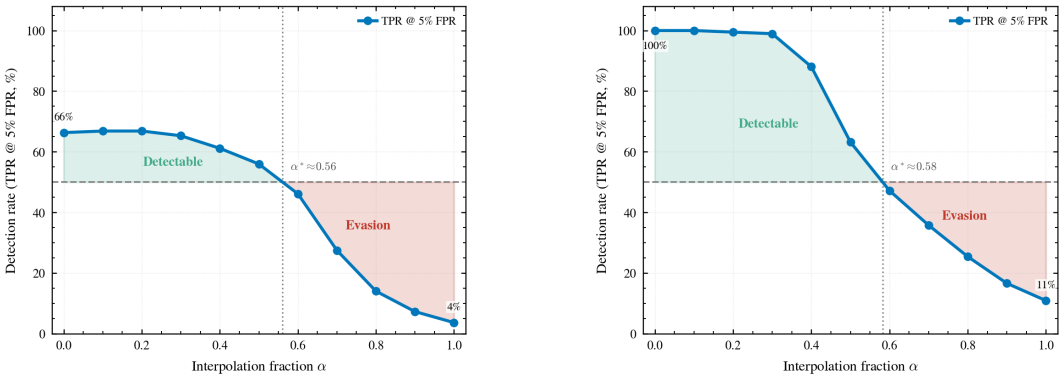
where c_{camo} is whichever of $\{c_{\text{atk}},\ \bar{c}_{\text{legit}}\}$ has ρ_1^{eff} closer to ρ_0 on that dataset (the harder camouflage), and c_{revealed} is the other. On XBlock $\rho_1^{\text{natural}} = 0.22 < \rho_0$, so $c_{\text{camo}} = \bar{c}_{\text{legit}}$ and $c_{\text{revealed}} = c_{\text{atk}}$; on Reddit $\rho_1^{\text{natural}} = 0.71 > \rho_0$, so the roles swap — $c_{\text{camo}} = c_{\text{atk}}$ and $c_{\text{revealed}} = \bar{c}_{\text{legit}}$. The sweep varies the attacker's effective correlation $\rho_1^{\text{eff}}(\alpha)$ continuously, and with it the measured gap $|\Delta\rho^{\text{eff}}(\alpha)| = |\rho_0 - \rho_1^{\text{eff}}(\alpha)|$, letting us read off the boundary at which the CB-RBM detection margin closes.

In Figures 8 and 9, α=1 denotes the hardest camouflage for that dataset (legit-mean context on XBlock, natural attack context on Reddit). This works because Relation Camouflage is a mirror operation in which the attacker picks whichever context distribution has ρ_1^{eff} closest to ρ_0, minimizing $|\Delta\rho^{\text{eff}}|$.

XBlock (ρ_0=0.30, ρ_1^{natural}=0.22). Legit context is *more* coherent than the attacker's own neighborhood, so injecting legit-mean context is the camouflage move. At α=0 TPR is 66% with $|\hat{\Phi} - \hat{\Psi}| \approx 320$; at α=0.5 TPR is 56%; at α=1 TPR collapses to 4% with $|\hat{\Phi} - \hat{\Psi}| \approx 181$; crossing at $\alpha^* \approx 0.56$.

Reddit (ρ_0=0.68, ρ_1^{natural}=0.71). The attacker's own context is already as coherent as the legit one, so the camouflage move is to keep attack context and *not* mix in legit-mean. At α=0 (pure legit-mean context) TPR is 100% with $\hat{\Phi}$ well below $\hat{\Psi}$; α=0.5 gives TPR 63%; at α=1 (pure attack context) TPR is 11% with $\hat{\Phi}$ crossing above $\hat{\Psi}$; crossing at $\alpha^* \approx 0.58$.

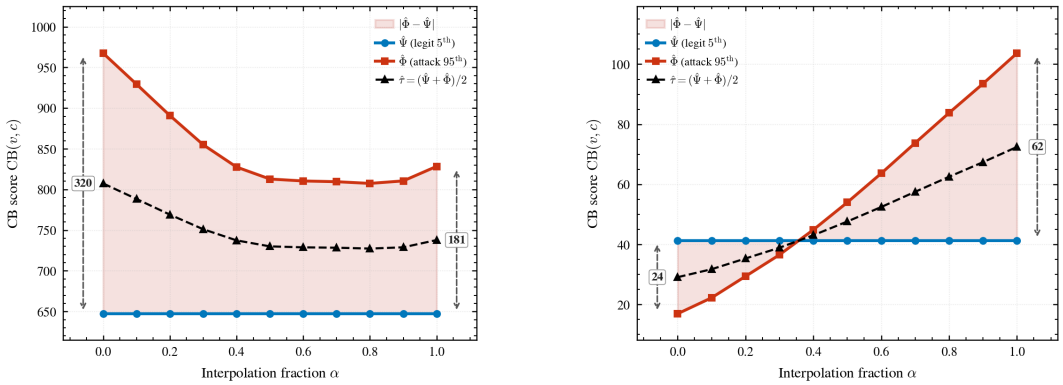
Consequence. The "hardest" camouflage is not a fixed recipe but the context distribution that minimizes $|\Delta\rho^{\text{eff}}|$ for the given dataset. For XBlock, legit-mean substitution minimizes $|\Delta\rho^{\text{eff}}|$; for



(a) XBlock: detection degrades as $\alpha \to 1$.

(b) Reddit: detection degrades as $\alpha \to 1$.

Figure 8: Context interpolation sweep, detection view. TPR at the calibrated 5%-FPR operating point vs α; the 50% crossing marks α^*.



(a) XBlock: $\hat{\Phi} \to \hat{\Psi}$ as $\alpha \to 1$.

(b) Reddit: $\hat{\Phi} \to \hat{\Psi}$ as $\alpha \to 1$.

Figure 9: Context interpolation sweep, separation view. Tail-band quantities $\hat{\Psi}$ (legit 5th-pctile), $\hat{\Phi}$ (attack 95th-pctile), and empirical tail-band midpoint $\hat{\tau} = (\hat{\Psi} + \hat{\Phi})/2$; shaded band is $|\hat{\Phi} - \hat{\Psi}|$, with magnitudes at α=0, 1 annotated.

Reddit, the attacker's own natural context already achieves this. In both regimes, CB-RBM's failure point is tied to a single scalar condition: the collapse of $|\Delta\rho^{\text{eff}}|$, consistent with the ρ_1-dependence of the Theorem 1 margin.

K Generalization across Domains

CB-RBM's two main-body datasets cover finance (XBlock) and social media (Reddit). This appendix tests whether the feature-context consistency principle transfers to further domains with different graph properties, label distributions, and fraud mechanics. Two additional organic datasets are reported: crowdsourcing (Tolokers) [55] and e-commerce review fraud (YelpChi) [56].

K.1 Tolokers (Crowdsourcing)

Tolokers [55] is a crowdsourcing graph with 11,758 worker nodes, 1,038,000 edges, 10 base features augmented to 19 with the same graph-structural signals used for Reddit (Appendix C), and a 21.8% unreliable-worker rate (organic labels). After sign correction $\rho_0 = 0.215$, placing it below XBlock (0.30) at the lower end of the homophily range we test, with a positive raw correlation gap. Tolokers is substantially denser than the main-body graphs (average degree 88 vs. Reddit's 15).

CB-RBM leads the attack benchmark on three of four attacks and places within 0.9 points of the top on Sybil (Table 18). Under 20-step Gaussian drift it retains a stability of 0.936 (FPR of 6.4%, near the

Table 18: Attack benchmark on Tolokers. Each cell: AUROC / TPR@5%FPR (%). Best per row in **bold**.

Attack	Energy-based		Supervised			Unsupervised		
	CB-RBM	Std-RBM	SAGE	GAT	GCN	DOM.	CoLA	GAE
Isolated Inj.	**96.4/89.3**	51.2/4.2	88.4/52.5	72.0/34.0	92.8/58.7	73.3/23.7	55.0/26.2	92.7/77.8
Feat. Camo	**97.5/87.8**	51.2/6.0	66.7/28.0	54.4/18.7	70.0/34.3	77.9/33.5	57.6/38.5	95.4/86.5
Sybil	98.9/96.7	51.2/4.2	97.1/86.7	98.9/**100**	99.1/**100**	97.3/96.7	68.1/40.5	**99.8/100**
Rel. Camo	**97.2/79.0**	51.2/4.2	79.4/13.7	80.1/22.3	77.6/15.7	68.6/15.5	63.8/11.7	57.6/12.7

Table 19: Tolokers drift evaluation at T=20: mean TPR (across the 4 attacks), drift stability, and Combined Score.

	Energy-based		Supervised			Unsupervised		
	CB-RBM	Std-RBM	SAGE	GAT	GCN	DOM.	CoLA	GAE
Mean TPR	**0.882**	0.047	0.452	0.438	0.522	0.424	0.292	0.693
Stability	0.936	0.676	0.890	0.934	**0.944**	0.940	0.782	0.686
Combined Score	**0.826**	0.031	0.403	0.409	0.493	0.398	0.229	0.475

calibrated 5% target), yielding a Combined Score of 0.826 — more than 0.33 above the best baseline (GCN at 0.493) and $\approx 27\times$ the Std-RBM ablation (Table 19). The detector's behavior on this denser, lower-ρ_0 domain is consistent with the feature-context consistency mechanism generalizing across feature modalities and homophily strengths where the raw correlation gap is positive.

K.2 YelpChi (E-commerce Review Fraud)

YelpChi is a review-fraud graph with $45{,}954$ review nodes, 14.5% labeled fraudulent reviews (organic), and three review-relationship edge types. We use the *net_rtr* edge subset (review-time-review, $1{,}147{,}232$ edges, average degree 25), which links reviews of the same product posted in the same month, the channel through which coordinated review rings surface. The other two edge types are either too sparse (*net_rur*: 42% of legitimate nodes have degree zero) or too dense for PyGOD baselines to finish in reasonable time (*net_rsr*: 6.8M edges). The 32 base features are augmented to 41 with the same nine graph-structural signals used for Reddit and Tolokers. After sign correction $\rho_0 = 0.293$; the natural fraud context has $\rho_1^{\text{natural}} = 0.306$, giving a raw correlation gap $\Delta\rho = -0.013$. This unfavorable gap indicates that coordinated fraud rings can retain neighborhood context *slightly more* coherent than that of typical legitimate reviews, so a compressed margin is expected, and the experiments below measure detection in this regime.

Table 20 shows the expected margin compression: CB-RBM's AUROC drops from the $\geq 96\%$ it holds on XBlock, Reddit, and Tolokers to the 74–79% range on three of the four attacks, and all three supervised GNNs — trained on YelpChi's abundant labels — overtake it on Combined Score (Table 21: SAGE 0.72, GCN 0.51, GAT 0.49 vs CB-RBM 0.41). Here the raw correlation gap is negative, and supervised baselines achieve higher Combined Scores. Two details confirm the mechanism rather than a generic performance loss. First, *Relation Camouflage* — the attack that most directly tests whether the context channel can distinguish rewired-to-legitimate-hubs adversaries from real legitimate nodes — is the only attack on which CB-RBM still leads on every metric: AUROC (98.2 vs SAGE's 97.5), TPR@5%FPR (86.5 vs 83.8), and Combined Score (0.81 vs 0.72); message-passing GNNs without the gated-context mechanism (GCN at 0.10 Combined, GAT at 0.37) suffer most, because the rewired context produces a coherence signature their attention or convolution cannot resolve. The feature-level and fabricated-context attacks (Iso, Feat, Sybil) rely on distinctions the compressed $\Delta\rho$ renders harder to separate, and supervised label access closes that gap. Second, CB-RBM remains the highest in drift stability of any method (0.933, narrowly above GCN's 0.930 and well above every other supervised and unsupervised baseline at ≤ 0.89) while Std-RBM collapses to near-chance everywhere and loses stability entirely (0.006 at T=20), confirming that context remains the essential signal even in this compressed-margin regime. YelpChi is therefore

Table 20: Attack benchmark on YelpChi (net_rtr). Each cell: AUROC / TPR@5%FPR (%). Best per row in **bold**.

Attack	Energy-based		Supervised			Unsupervised		
	CB-RBM	Std-RBM	SAGE	GAT	GCN	DOM.	CoLA	GAE
Isolated Inj.	76.6/34.5	50.3/10.5	**98.6/93.3**	79.0/53.0	87.2/74.3	82.7/26.5	80.3/35.0	73.6/20.5
Feat. Camo	78.9/41.5	62.9/14.5	**89.6/62.7**	56.8/27.0	64.4/35.8	80.3/15.0	85.5/48.5	70.0/12.5
Sybil	74.9/15.0	50.3/10.5	98.9/94.2	99.5/**100.0**	**99.8/100.0**	73.1/15.5	60.0/5.5	61.7/12.0
Rel. Camo	**98.2/86.5**	50.3/10.5	97.5/83.8	82.7/41.7	79.2/11.0	72.8/15.5	61.8/11.5	61.5/11.5

Table 21: YelpChi (net_rtr) drift evaluation at $T{=}20$: mean TPR (across the 4 attacks), drift stability, and Combined Score.

	Energy-based		Supervised			Unsupervised		
	CB-RBM	Std-RBM	SAGE	GAT	GCN	DOM.	CoLA	GAE
Mean TPR	0.444	0.115	**0.835**	0.554	0.553	0.181	0.251	0.141
Stability	**0.933**	0.006	0.861	0.885	0.930	0.598	0.879	0.443
Combined Score	0.414	0.001	**0.718**	0.490	0.514	0.108	0.221	0.062

an empirical boundary case. When the raw correlation gap is unfavorable, the supervised methods' label-access advantage catches up, though only on the attacks that do not specifically exercise the context-coherence signal.

L CB-RBM as a Drift-Filter Defense

Theorem 1 certifies separation only when the positive gated-context activation that adversarial context can produce stays below the legitimate dominance margin (Assumptions 4 and 5). On YelpChi the raw correlation gap is unfavorable ($\Delta\rho = -0.013$), and supervised baselines accordingly outperform CB-RBM standalone (Appendix K.2). This appendix tests whether CB-RBM still adds operational value in this regime. The setting we evaluate: an operator runs a deployed primary detector — typically a supervised model trained on labeled adversaries, or one pinned by operational constraints — whose false-alarm rate inflates under drift but which cannot easily be retrained. Layered downstream, CB-RBM suppresses alarms whose flagged nodes remain within the legitimate distribution learned at calibration, recovering precision without retraining or new labels. The *rescue rate* measures the fraction of drift-induced false positives that CB-RBM correctly suppresses.

Across all seven baselines, CB-RBM rescues 87.6–93.5% of drift-induced false positives (Table 22). Even where the raw correlation gap is unfavorable and CB-RBM is not the top standalone detector, layering it as a downstream filter preserves the method's operational value in this empirical regime. CB-RBM is an additive layer with no retraining cost, no label cost, and no architectural integration — a drop-in complement in such scenarios, not a replacement.

Established adversaries. Scoring naturally occurring adversaries with their accrued neighborhood context gives AUROC 0.84 on XBlock and 0.54 on Reddit. At $T{=}20$ on XBlock, using CB-RBM as a downstream filter reduces DOMINANT's drift FPR from 98.3% to 6.5%.

M Transferability of the Context Boost Principle

The Context Boost principle — score a node by the *differential* of a learned density under the presence vs absence of its neighborhood context — is not specific to the RBM. CB-RBM's algebraic cancellation in Eq. (4) (the quadratic feature term drops out of the subtraction) makes the construction particularly clean, but any generative model whose likelihood decomposes additively between a context-free and a context-conditional branch admits an analogous score. This appendix outlines the

Table 22: YelpChi rescue rates (%) at T=20.

	Unsupervised				Supervised		
	Std-RBM	DOMINANT	GAE	CoLA	SAGE	GAT	GCN
YelpChi	93.5	89.9	93.2	87.6	91.0	90.9	92.9

structural requirements for porting the principle to other architectures and discusses what transfers automatically vs what needs re-derivation.

Structural requisites. To instantiate the Context Boost principle on a host model, three properties must hold. *(i) Decomposable scalar score.* A function $S(\mathbf{v}, \mathbf{c})$ on the joint input that can also be evaluated with context ablated, so that $\mathrm{CB}(\mathbf{v}, \mathbf{c}) = S(\mathbf{v}, \mathbf{c}) - S(\mathbf{v}, \varnothing)$ is well-defined and captures the incremental information content of $\mathbf{c}$. *(ii) Gated context pathway.* A mechanism that selectively suppresses uninformative or noisy context dimensions, so that random or absent context produces near-zero CB; CB-RBM uses a GLU gate, but attention, gating units, or learnable context masking all serve. *(iii) Lipschitz continuity of S.* A bounded Lipschitz constant in both channels is necessary for Theorem 2's Wasserstein drift-stability argument to port.

Candidate instantiations. Several generative families satisfy these requisites and admit a natural Context Boost analogue. A *variational autoencoder* with a context-conditional encoder yields $\mathrm{CB}_{\mathrm{VAE}} = \mathrm{ELBO}(\mathbf{v}, \mathbf{c}) - \mathrm{ELBO}(\mathbf{v}, \varnothing)$, with context-attention playing the gating role. A *normalizing flow* gives $\mathrm{CB}_{\mathrm{NF}} = \log p(\mathbf{c} \mid \mathbf{v})$ directly from invertibility, and Lipschitz-constrained flows [57] make Theorem 2 immediately applicable. A *masked-modeling transformer* over concatenated $(\mathbf{v}, \mathbf{c})$ tokens produces CB_{T} as the gap between context-present and context-dropout NLL, with cross-attention weights on context tokens supplying the gate. In each case the operational payoff is the same: feature-context consistency is captured by a single decomposable score that supports detection and inherits drift stability via Lipschitz continuity — both from within a single model.

What transfers automatically, what does not. Assumptions 1 (signed structural correlation) and 3 (bounded moments) are properties of the data, independent of host architecture — any model inherits them. Assumptions 2 (coherent activation), 4 (adversarial context gap) and 5 (coherent dominance) are stated on the trained context pathway, so they and the constants that appear in the separation and drift bounds ($\beta_0, p_0, \gamma, L_v, L_c$) must be re-verified or reformulated for the host model. The per-unit decomposition Δ_j used in the proof of Theorem 1 is also CB-RBM–specific; an analogous decomposition into architecture-native primitives (latent units for VAE, flow layers for normalizing flows, attention heads for transformers) is required to recover the $\Omega(n_h)$-style separation scaling. What universally carries over is the framework: feature-context consistency as a drift-stable detection signal, operationalized as a decomposable score on a gated context pathway. CB-RBM is one convenient instantiation; the principle itself is broader.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: All numerical claims (AUROC ≥ 0.96 on four attacks, Combined Score ≥ 0.87 at T=20, calibrated 5% FPR, Std-RBM FPR collapse to 89–100%) are directly supported by Tables 1, 3 and Figures 1, 3. The zero-label claim is enforced by construction (threshold set from legitimate data only). The theory (Assumptions 1–5, Theorems 1–2) gives conditional worst-case results, with numerical comparisons of the bounds at root-mean-square constants in Appendix I; detection and drift performance are measured directly in §6.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: §6.5 ("Limitations") identifies six scope boundaries: detection relies on $\rho_0 > 0$, which can fail in random or adversarially curated topologies; detection degrades as $\Delta\rho$ collapses, where supervised baselines with adequate labels can perform better; adversaries who acquire genuine hub connections (Relation Camouflage) reduce the detection gap and can become indistinguishable from legitimate users in the limit; substantial topology change can impair calibration at the frozen threshold and needs the label-free threshold refresh; white-box adversaries with gradient access can lower detection; and the drift analysis assumes smooth drift, so abrupt or adversarial shifts may need additional mechanisms. Transferability across architectures is discussed in Appendix M; boundary-regime generalization in Appendix K.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Assumptions 1–5 are stated in §5 with informal justifications, empirical verifiability conditions, and remarks on how s-correction ensures $\rho_0 > 0$. Measured values: $\rho_0 = 0.30$ (XBlock) and $\rho_0 = 0.68$ (Reddit); coherent-activation fraction β_0 and bounded-moment constants M_x, M_c verified after training on both datasets. Proof sketches appear in §5; complete proofs with supporting lemmas in Appendix B.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: §6.1 specifies all hyperparameters (n_h=256 for XBlock, n_h=128 for Reddit, CD-1, λ_{sp}=0, 100 epochs) and evaluation protocol (95th-percentile of $-\text{CB}$ over legitimate nodes, TPR@5%FPR metric). Dataset details: XBlock (153K nodes, 38 features, ρ_0=0.30) and Reddit (11K nodes, 73 features, ρ_0=0.68), both publicly available. Attack constructions are defined in §3, with the context-interpolation sweep in Appendix J. The six GADBench baselines use GADBench defaults [46], and the nine further detectors follow their published configurations with model selection on a held-out validation split (Appendix D).

5. **Open access to data and code**

   Answer: [Yes]

   Justification: The reference implementation of CB-RBM, the four attacks and the drift simulation is available at `https://github.com/serval-uni-lu/CB-RBM` under the MIT license. All four datasets used in this work (XBlock [21], Reddit [22], Tolokers [55], YelpChi [56]) are publicly available, and the feature construction is described in Appendix C.

6. **Experimental setting/details**

Answer: [Yes]

Justification: §6.1 specifies all training details. Appendix C provides feature categories and counts for each dataset, sign-vector construction, normalization, and context computation; full feature definitions are in the released code. Appendix G reports hyperparameter sensitivity (n_h, pooling operator).

7. **Experiment statistical significance**

   Answer: [Yes]

   Justification: The attack benchmark reports means over 3 random seeds with standard deviations in §6.3; XGBGraph, Graph Transformer, BWGNN and GDN use 5 seeds (§6.4, Appendix D); the sign-vector experiment reports standard deviations over 5 seeds (Appendix E.2) and the threshold-drift experiment over 3 seeds (Appendix I.2).

8. **Experiments compute resources**

   Answer: [Yes]

   Justification: All experiments run on a single GPU. CB-RBM training takes under 10 minutes per dataset, and the full benchmark, drift and ablation suite completes within a standard single-GPU budget. CONAD and AnomalyDAE exceeded memory on XBlock and are reported on Reddit only (Appendix D).

9. **Code of ethics**

   Answer: [Yes]

   Justification: All datasets used are publicly available and anonymized: XBlock (anonymized Ethereum wallet addresses), Reddit (anonymized user identifiers), Tolokers (anonymous worker IDs), YelpChi (anonymized reviewer identifiers). The work aims to improve fraud detection, a socially beneficial application. False-positive rates on legitimate users under drift are reported in §6.4 and Appendix I.

10. **Broader impacts**

    Answer: [Yes]

    Justification: §1 motivates the work by the scale of fraud losses. §6.5 ("Limitations") states where the detector fails, namely graphs without feature-context correlation, adversaries with genuine connections, topology change and white-box adversaries. False alarms on legitimate accounts are the main negative consideration, and the paper reports them under drift at the calibrated threshold (§6.4). In deployment, τ^* is an operator-set false-positive budget (5% here), the flagged share of legitimate nodes stays at or below 16% under the five drift regimes of Table 11, and the score should rank cases for human review rather than block accounts automatically, since new or sparsely connected legitimate accounts can lack neighborhood context.

11. **Safeguards**

    Answer: [N/A]

    Justification: The paper proposes a fraud detection method, not a generative model or dataset release with misuse potential. The attack constructions described are standard synthetic evaluation tools, not deployable attack tooling.

12. **Licenses for existing assets**

    Answer: [Yes]

    Justification: XBlock [21], Reddit [22], Tolokers [55], and YelpChi [56] are all publicly available and properly cited. All are standard academic benchmarks used under their published terms.

13. **New assets**

    Answer: [N/A]

    Justification: No new datasets are released. The code is released under the MIT license at https://github.com/serval-uni-lu/CB-RBM.

14. **Crowdsourcing and research with human subjects**

    Answer: [N/A]

Justification: No crowdsourcing or human subjects research. The Tolokers dataset consists of pre-existing, anonymized crowdsourcing graph data from prior work [55].

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Answer: [N/A]

    Justification: No human subjects research.

16. **Declaration of LLM usage**

    Answer: [N/A]

    Justification: LLMs are not used as a component of the core method or experiments. Any use for writing assistance does not impact the scientific content or experimental results.