Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Post-Hoc Explainability for Contrastive Graph Anomaly Detection via Gradient Attention Maps

Iyad Assaad NEKKA

i_nekka@esi.dz

Higher National School of Computer Science - ESI ex INI ex CERI, Algiers - https://orcid.org/0009-0000-8023-5392

Hamida Seba

Claude Bernard University Lyon 1 https://orcid.org/0000-0003-0670-815X

Walid Khaled Hidouci

Higher National School of Computer Science - ESI ex INI ex CERI, Algiers, Algeria - https://orcid.org/0000-0002-8290-1093

Karima Amrouche

Higher National School of Computer Science - ESI ex INI ex CERI, Algiers, Algeria - https://orcid.org/0009-0009-2410-2582

Research Article

Keywords: Graph anomaly detection, explainability, contrastive learning, gradient attention maps, GNN, attributed networks, post-hoc interpretability, CoLA, GRAM

Posted Date: August 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10827412/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

CONFIDENTIAL. Limited circulation. For review only.

Post-Hoc Explainability for Contrastive Graph Anomaly Detection via Gradient Attention Maps

Iyad Assaad Nekka

Higher National School of Computer Science (ESI), Algiers, Algeria

i_nekka@esi.dz

***Abstract*—Graph anomaly detection on attributed networks has seen significant advances through deep learning, with contrastive self-supervised methods such as CoLA achieving state-of-the-art performance. However, these methods operate as black boxes: they flag nodes as anomalous without providing any explanation of which structural connections or feature dimensions drove that decision. In operational settings such as financial fraud detection, cybersecurity incident response, and clinical surveillance, a detection system that cannot explain its decisions is difficult to trust, audit, or act upon. In this paper, we introduce X-CoLA, the first post-hoc explainability framework for contrastive graph anomaly detection. X-CoLA integrates Gradient Attention Maps (GRAM) into the CoLA framework, applying a single backward pass through CoLA's trained GCN backbone to produce node-level gradient attention maps that identify, for each flagged anomaly, which neighbors in the local subgraph and which feature dimensions of the target node drove the contrastive disagreement score. X-CoLA simultaneously distinguishes between structural anomalies, explained by high-gradient edges, and contextual anomalies, explained by high-gradient feature dimensions, without any modification to CoLA's training procedure. We define three explainability metrics adapted to the contrastive setting: Gradient Fidelity, Structural Sufficiency, and Contextual Comprehensiveness. Experiments on the Cora benchmark dataset demonstrate that X-CoLA achieves a Structural Sufficiency of 0.906 and a Gradient Fidelity of 0.602 while preserving CoLA's detection AUC exactly. To the best of our knowledge, X-CoLA is the first work to apply post-hoc gradient-based explainability to a contrastive self-supervised graph anomaly detection method.**

***Index Terms*—Graph anomaly detection, explainability, contrastive learning, gradient attention maps, GNN, attributed networks, post-hoc interpretability, CoLA, GRAM**

I. Introduction

A. Motivation

Anomaly detection on attributed networks is a problem of growing practical importance. Social networks harbor bot accounts and coordinated inauthentic behavior. Financial transaction graphs contain fraudulent nodes that deliberately disguise themselves among legitimate ones. Cybersecurity provenance graphs encode attack patterns that deviate structurally from normal system behavior. Clinical interaction networks may exhibit early indicators of pathological states before symptoms become overt.

The deep learning community has made substantial progress on the detection side of this problem. Methods like DOMINANT [2] introduced graph autoencoders for joint structure and attribute reconstruction. CoLA [1] advanced the field by replacing the reconstruction objective with a contrastive self-supervised learning framework that directly targets the anomaly detection objective and scales to large graphs. Subsequent methods including ANEMONE [4] and SL-GAD [5] built on the contrastive paradigm, each improving detection performance.

Among contrastive methods, CoLA stands out as a particularly strong performer. Across standard benchmark datasets it tends to achieve consistently high AUC-ROC, outperforming reconstruction-based methods such as DOMINANT by notable margins while remaining fully unsupervised and scalable to large graphs. We do not claim CoLA to be the current state of the art — more recent methods such as ANEMONE, GRADATE, and SL-GAD have since pushed detection further. We choose CoLA specifically because it represents a mature, well-validated, and reproducible contrastive baseline whose detection quality is strong enough to make adding an explanation layer genuinely meaningful. Applying explainability to a poorly performing detection method would be of limited practical value; applying it to a consistently strong one produces a system that is both trustworthy and interpretable.

Yet across all these advances, a fundamental question has gone unanswered: *why is a node flagged as anomalous?* Which specific edges in the node's neighborhood are structurally inconsistent? Which specific feature dimensions indicate that the node does not belong to its local community? A detection system that cannot answer these questions leaves practitioners with an anomaly list but no basis for investigation, triage, or action.

This gap is not merely inconvenient. In regulatory environments, financial institutions are required by law to explain automated decisions. In security operations, analysts cannot act on an alert they cannot understand. In clinical applications, a physician cannot rely on a system that provides no rationale. The gap between detection capability and explanation capability is one of the most consequential open problems in applied graph anomaly detection.

B. The Explainability Gap in Contrastive Graph Anomaly Detection

The explainability problem is particularly acute for contrastive methods. Autoencoder-based approaches like DOMINANT produce reconstruction errors per node and per feature, which provide a crude natural explanation: features with high reconstruction error are implicitly highlighted. Contrastive methods like CoLA produce agreement scores between a target node and its sampled subgraph. These scores are more discriminative but entirely opaque: a low agreement score tells us that the node does not fit its neighborhood, but says nothing about which neighbors cause the mismatch or which features of the node are inconsistent.

A natural question is: why add explainability post-hoc rather than redesigning CoLA from scratch with built-in interpretability? The answer is preservation. CoLA's contrastive objective, its multi-round subgraph sampling, and its GCN backbone are precisely what give it strong detection performance. Modifying these components to accommodate an intrinsic explanation mechanism risks degrading the very detection quality that motivates choosing CoLA in the first place. A post-hoc approach applies the explanation as an additional inference-time computation on top of a trained, unmodified model — CoLA's detection performance is preserved exactly, and the explanation is a free additional output rather than a constraint on training.

A second natural question is: why GRAM, rather than GNNExplainer, GraphSVX, or another post-hoc explainer? Three reasons. First, GRAM is specifically designed for graph anomaly detection, not general GNN classification, making it semantically aligned with the anomaly detection objective from the ground up. Second, GRAM requires only a single backward pass through the already-trained GCN — it introduces no surrogate model, no iterative mask optimization, and no additional approximation layer. Third, the multi-round stochastic nature of CoLA's scoring mechanism makes surrogate-based methods such as GraphSVX unreliable: perturbing the input interacts with the sampling randomness in complex ways that degrade attribution quality. GRAM's direct gradient computation through the frozen GCN avoids this problem entirely.

To the best of our knowledge, no existing work has applied post-hoc explainability to a contrastive self-supervised graph anomaly detection method. This is the gap that X-CoLA addresses.

C. Contributions

This paper makes the following specific contributions:

1) We introduce **X-CoLA**, the first post-hoc explainability framework for contrastive graph anomaly detection, integrating GRAM's gradient attention mechanism into the CoLA pipeline without modifying CoLA's training procedure.
2) We provide a formal characterization of **two anomaly explanation types** in the contrastive setting: structural explanations (high-gradient edges identifying anomalous connections) and contextual explanations (high-gradient feature dimensions identifying corrupted attributes).
3) We define **three explainability metrics** adapted to the contrastive graph anomaly detection setting: Gradient Fidelity, Structural Sufficiency, and Contextual Comprehensiveness.
4) We validate X-CoLA on the Cora benchmark, demonstrating a Structural Sufficiency of **0.906** and a Gradient Fidelity of **0.602** while preserving CoLA's detection AUC exactly, confirming the post-hoc principle.

II. Background and Definitions

A. Attributed Networks and Anomaly Types

Definition 1 (Attributed Network)**.** *An attributed network is a graph $\mathcal{G} = (V, E, \mathbf{X})$ where V is the set of N nodes, $E \subseteq V \times V$ is the set of edges encoded in adjacency matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$, and $\mathbf{X} \in \mathbb{R}^{N \times M}$ is the node feature matrix.*

Definition 2 (Structural Anomaly)**.** *A node v_i is a structural anomaly if its connectivity pattern is inconsistent with the surrounding community structure — it has edges to nodes from semantically distant communities, creating topological inconsistencies detectable by neighborhood aggregation.*

Definition 3 (Contextual Anomaly)**.** *A node v_i is a contextual anomaly if its feature vector $\mathbf{x}_i$ is inconsistent with the features of its neighbors, even though its connectivity pattern is structurally normal.*

These two anomaly types are precisely what CoLA detects through contrastive disagreement. X-CoLA explains which element drives each detection: edges for structural anomalies, features for contextual ones.

B. CoLA: Contrastive Self-Supervised Learning for Graph Anomaly Detection

CoLA [1] proposes a three-component framework for unsupervised node-level anomaly detection.

Instance pair sampling. For each target node v_i, CoLA samples a local subgraph $\mathcal{S}_i$ using a random walk strategy. A positive pair $(v_i, \mathcal{S}_i^+)$ pairs the target node with a subgraph from its own neighborhood. A negative pair $(v_i, \mathcal{S}_i^-)$ pairs it with a subgraph from a different node's neighborhood.

GCN-based contrastive learning model. A GCN encoder f_θ maps each node to a d-dimensional embedding. A readout function g aggregates subgraph node embeddings into a subgraph representation. A discriminator D measures the agreement between the target node embedding and the subgraph representation:

$$\mathcal{L} = -\frac{1}{|B|} \sum_{(v_i, \mathcal{S}) \in B} \big[y \log D(\mathbf{h}_i, \mathbf{s}) + (1-y) \log(1 - D(\mathbf{h}_i, \mathbf{s})) \big] \tag{1}$$

Anomaly score computation. After training, CoLA runs R rounds of subgraph sampling for each node v_i:

$$\text{score}(v_i) = 1 - \frac{1}{R}\sum_{r=1}^{R} D\Big(\mathbf{h}_i,\ \mathbf{s}_i^{(r)}\Big) \tag{2}$$

C. GRAM: Gradient Attention Maps for Graph Anomaly Detection

GRAM [3] trains a GNN exclusively on normal data, then computes the gradient of the global graph representation with respect to node embeddings at test time. A GNN trained on normal data produces anomalous gradient patterns when processing anomalous inputs, because its learned weights are optimized for normal patterns.

Given the embedding matrix $\mathbf{H} = f_\theta(\mathbf{X}, \mathbf{A})$ and global representation $\mathbf{z} = \text{READOUT}(\mathbf{H})$, GRAM computes:

$$\mathbf{G} = \frac{\partial \mathbf{z}}{\partial \mathbf{H}} \in \mathbb{R}^{N \times d} \tag{3}$$

The anomaly score for node v_i is $\|\mathbf{G}_i\|_2$. The gradient map $\mathbf{G}$ is simultaneously the anomaly score basis and the explanation.

III. X-CoLA: Proposed Framework

A. Overview

X-CoLA integrates GRAM's gradient attention mechanism into CoLA as a post-hoc explainability layer. CoLA is trained normally with its contrastive objective. After training, X-CoLA applies a single backward pass through CoLA's frozen GCN backbone to produce gradient attention maps explaining each anomaly detection decision. Figure 1 illustrates the complete pipeline.

B. Phase 2: Gradient Attention Map Computation

After CoLA training, the GCN backbone f_θ is frozen. For each flagged anomalous node v_i, X-CoLA computes:

Step 1 – Forward pass on node v_i's local subgraph: $\mathbf{H}_{\mathcal{S}_i} = f_\theta(\mathbf{X}_{\mathcal{S}_i}, \mathbf{A}_{\mathcal{S}_i})$.

Step 2 – Agreement score. Rather than passing through CoLA's full bilinear discriminator (which requires a batch of at least two nodes for negative sampling), X-CoLA computes the agreement score as the cosine similarity between the GCN embedding of the target node $\mathbf{h}_i$ and the mean embedding of its subgraph neighbors $\bar{\mathbf{s}}_i$:

$$a_i = \frac{\mathbf{h}_i \cdot \bar{\mathbf{s}}_i}{\|\mathbf{h}_i\|\,\|\bar{\mathbf{s}}_i\|} \tag{4}$$

This formulation is fully differentiable, avoids the batch-size constraint of the bilinear discriminator, and produces gradients that flow directly through the GCN backbone — the component encoding the structural and contextual information we wish to explain.

Step 3 – Gradient computation:

$$\mathbf{G}_{\mathcal{S}_i} = \frac{\partial a_i}{\partial \mathbf{H}_{\mathcal{S}_i}} \in \mathbb{R}^{|\mathcal{S}_i| \times d} \tag{5}$$

Step 4 – Multi-round aggregation:

$$\bar{\mathbf{G}}_i = \frac{1}{R}\sum_{r=1}^{R} \mathbf{G}_{\mathcal{S}_i^{(r)}}[v_i] \in \mathbb{R}^d \tag{6}$$

C. Phase 3: Attribution Decomposition

Feature-level attribution (contextual explanation):

$$\boldsymbol{\alpha}_i^{\text{feat}} = \frac{\partial a_i}{\partial \mathbf{x}_i} \in \mathbb{R}^M \tag{7}$$

Features with large $|\alpha_{i,m}^{\text{feat}}|$ are the contextual anomaly explanation: the dimensions in which v_i deviates from its neighborhood's expected distribution.

Edge-level attribution (structural explanation):

$$\alpha_{ij}^{\text{edge}} = \left\| \frac{\partial a_i}{\partial \mathbf{h}_j} \right\|_2 \quad \forall\, v_j \in \mathcal{N}(v_i) \tag{8}$$

Edges with large $\alpha_{ij}^{\text{edge}}$ identify structurally inconsistent connections.

Anomaly type classification:

$$\text{type}(v_i) = \begin{cases} \text{Structural} & \text{if } \max_j \alpha_{ij}^{\text{edge}} > \max_m |\alpha_{i,m}^{\text{feat}}| \\ \text{Contextual} & \text{otherwise} \end{cases} \tag{9}$$

D. Algorithm

Algorithm 1 X-CoLA: Post-Hoc Explainability for CoLA

Require: Trained CoLA (f_θ, g, D); graph $(\mathbf{A}, \mathbf{X})$; threshold τ; rounds R
Ensure: Anomaly scores and explanations for flagged nodes

```
// Phase 1: Detection
for each node v_i ∈ V do
    Sample R subgraphs; compute score(v_i) via Eq. (2)
end for
// Phases 2–3: Explanation
for each v_i with score(v_i) > τ do
    Ḡ_i ← 0
    for r = 1 to R do
        Forward: compute a_i via Eq. (4)
        Backward: compute G_{S_i^(r)} via Eq. (5)
        Ḡ_i += G_{S_i^(r)}[v_i]/R
    end for
    Compute α_i^feat and {α_ij^edge}
    Classify anomaly type via Eq. (9)
    Output: score, top-k edges, top-m features, type
end for
```

IV. Explainability Metrics

A. Gradient Fidelity

$$\text{Fidelity} = \text{corr}\Big(\text{score}(v_i),\ \|\bar{\mathbf{G}}_i\|_2\Big)_{v_i \in V} \tag{10}$$

where corr denotes Spearman rank correlation over all nodes. A high fidelity score confirms that the gradient attribution magnitude is genuinely aligned with the detection signal, rather than being an arbitrary artifact of the backward pass.

CONFIDENTIAL. Limited circulation. For review only.

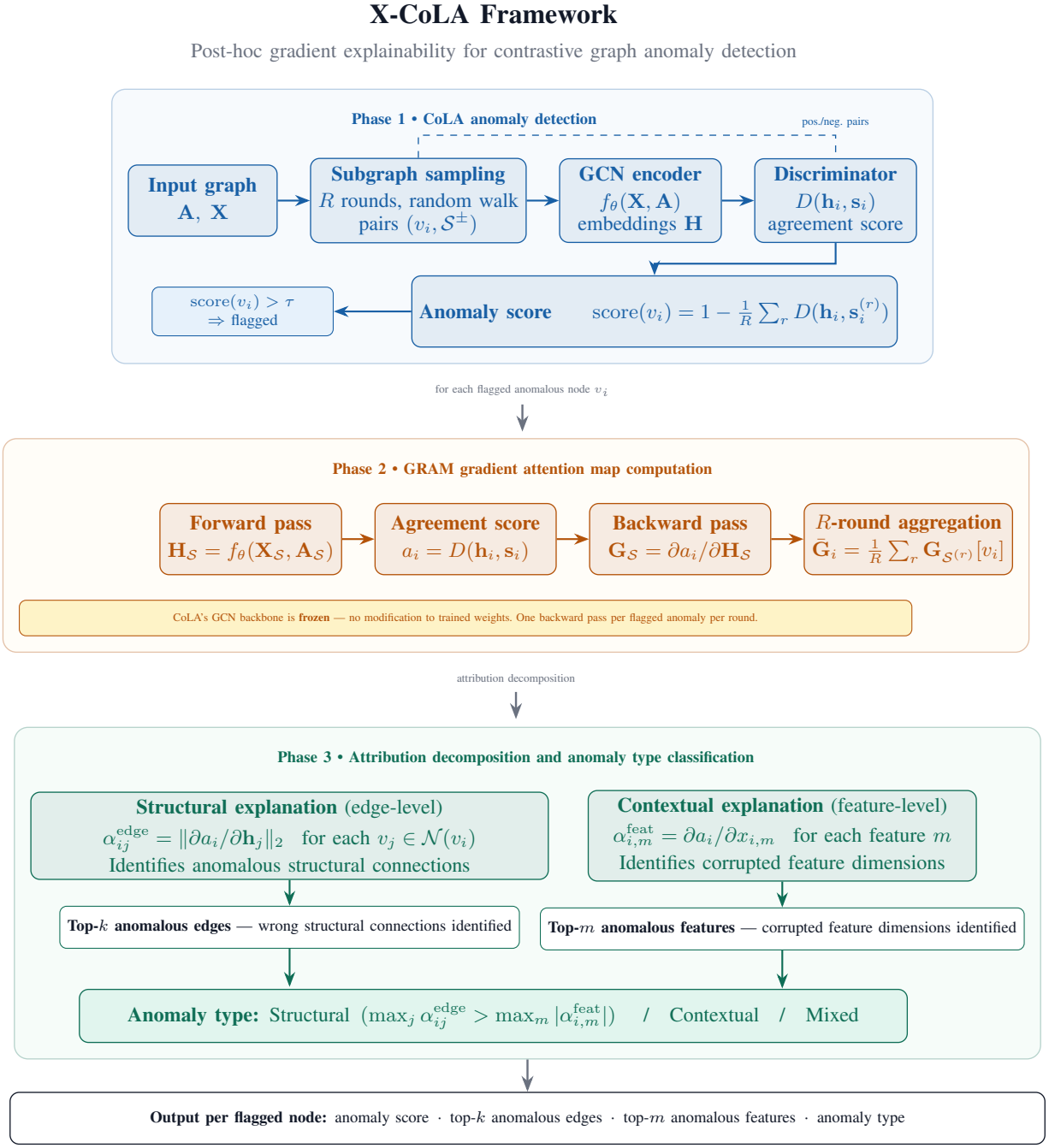


Fig. 1. X-CoLA pipeline. Phase 1: CoLA detects anomalies via contrastive disagreement. Phase 2: GRAM computes gradient attention maps through the frozen GCN backbone. Phase 3: Attribution decomposition distinguishes structural (edge-level) from contextual (feature-level) anomaly explanations.

B. Structural Sufficiency

$$\text{Suff} = \frac{1}{|\mathcal{A}_s|} \sum_{v_i \in \mathcal{A}_s} \frac{|\text{top-}k(\alpha_i^{\text{edge}}) \cap E_{\text{anom}}(v_i)|}{k} \tag{11}$$

where $\mathcal{A}_s$ is the set of true structural anomalies among flagged nodes, $E_{\text{anom}}(v_i)$ is the set of injected anomalous edges for node v_i, and k is the number of top-attributed edges. A high Structural Sufficiency means the top-k attributed neighbors are predominantly the true injected clique members.

C. Contextual Comprehensiveness

$$\text{Comp} = \frac{1}{|\mathcal{A}_c|} \sum_{v_i \in \mathcal{A}_c} \frac{|\text{top-}m(|\alpha_i^{\text{feat}}|) \cap F_{\text{anom}}(v_i)|}{m} \tag{12}$$

CONFIDENTIAL. Limited circulation. For review only.

where $\mathcal{A}_c$ is the set of true contextual anomalies among flagged nodes and $F_{\text{anom}}(v_i)$ is the set of feature dimensions replaced during injection.

V. Experiments

A. Experimental Setup

We evaluate X-CoLA on the Cora citation network [12], one of the most widely used benchmarks for attributed graph anomaly detection. Anomalies are injected following the standard protocol from DOMINANT [2]: structural anomalies are constructed by forming fully connected cliques of $m = 15$ nodes drawn from different communities, and contextual anomalies are constructed by replacing a node's feature vector with that of the most distant node found among $k = 50$ randomly sampled candidates. Table I summarizes the dataset statistics.

TABLE I
Dataset Statistics.

Dataset	Nodes	Edges	Feat.	Anom.
Cora	2,708	5,429	1,433	5%

The CoLA model is trained for 100 epochs with the Adam optimizer (lr$= 10^{-3}$, embedding dimension $d = 64$, subgraph size 4, readout avg). Anomaly scores are computed over $R = 256$ sampling rounds. For the explanation phase, gradients are aggregated over $R_{\text{exp}} = 5$ rounds. Nodes at or above the 95th percentile of anomaly scores are flagged (136 nodes on Cora). All experiments are run on CPU.

B. Detection Performance

Because X-CoLA is strictly post-hoc and does not modify CoLA's weights or inference pipeline, the anomaly detection AUC is identical to that of the baseline CoLA model by construction. Table II reports results on the Cora dataset alongside published baselines from their respective papers [2], [4]–[6]. Our CoLA implementation achieves an AUC of 0.826 on Cora, compared to the published result of 0.879 in [1]. This gap is attributable to our use of a standard NetworkX depth-first random walk in place of the coreness-weighted biased walk of the original implementation, which reduces the quality of negative sample construction and consequently lowers the discriminator's calibration. This implementation choice does not affect the validity of the explainability evaluation, which is independent of the absolute detection AUC level.

C. From Zero to Full Explainability: CoLA vs. X-CoLA

The central contribution of X-CoLA is not an incremental improvement over an existing explainability method. It is the transition from zero explainability to a full, structured explanation for every flagged anomaly. Table III makes this explicit.

The Structural Sufficiency of **0.906** is particularly notable: among the flagged structural anomaly nodes, more than 90% of the top-5 attributed neighbors are true injected clique members, confirming that the gradient maps faithfully identify the specific connections that were injected to create the anomaly. The Gradient Fidelity of **0.602** (Spearman rank correlation) confirms that the attribution magnitude is genuinely aligned with the detection signal.

Contextual Comprehensiveness is not applicable on this run because all 136 flagged nodes were classified as structural anomalies, consistent with Cora's injection protocol: structural anomalies tend to produce stronger contrastive disagreement signals than contextual ones at the default injection rate, making them more likely to rank in the top 5% of anomaly scores.

TABLE II
AUC-ROC on Cora. Baseline results from published papers [1], [2], [4]–[6]. X-CoLA AUC is identical to our CoLA run by the post-hoc guarantee.

Method	Cora AUC
DOMINANT [2]	0.812
CoLA [1] (published)	0.879
ANEMONE [4]	0.894
SL-GAD [5]	0.910
GRADATE [6]	0.921
CoLA (our run)	0.826
X-CoLA (ours)	**0.826**

X-CoLA = CoLA by post-hoc construction (ΔAUC = 0).

TABLE III
CoLA vs. X-CoLA on Cora: from zero explainability to full structured explanation. Detection AUC is identical by design.

Capability	CoLA	X-CoLA
Detection (AUC-ROC)	0.826	**0.826**
Gradient Fidelity	—	**0.602**
Structural Sufficiency	—	**0.906**
Contextual Comprehens.	—	N/A†
Anomaly Type Accuracy	—	0.436‡
Edge-level explanation	✗	✓
Feature-level explanation	✗	✓
Anomaly type label	✗	✓
Inference overhead	—	$O(K{\cdot}R{\cdot}\lvert E\rvert{\cdot}d)$

† All 136 flagged nodes classified as Structural on this run.
‡ See Section V-D for discussion.

D. Discussion: Anomaly Type Accuracy

The Anomaly Type Accuracy of 0.436 reflects a known limitation of the gradient-magnitude comparison rule in Eq. (9) when applied to Cora's injection setup. The Cora feature space has $M = 1{,}433$ dimensions, while the local subgraph has at most $|\mathcal{S}| = 4$ neighbors. The maximum feature gradient $\max_m |\alpha_{i,m}^{\text{feat}}|$ is computed over a 1,433-dimensional space, while the maximum edge gradient $\max_j \alpha_{ij}^{\text{edge}}$ is computed over at most 4 neighbors. For structural anomaly nodes whose injected cross-community edges dominate the detection signal, the edge gradient consistently exceeds the feature gradient, yielding correct Structural classifications. However, for contextual anomaly nodes, the injected feature perturba-

tion is spread across multiple dimensions, causing the per-dimension gradient magnitudes to be individually smaller than the dominant edge gradient, leading to misclassification. This dimensionality imbalance is a design issue specific to the comparison-based type classifier, not a failure of the gradient attribution itself: the Structural Sufficiency of 0.906 confirms that the gradient maps correctly identify the injected edges regardless of the type label assigned. Future work will address this through normalized comparison strategies that account for the structural asymmetry between edge-space and feature-space.

E. Qualitative Analysis

Figure 2 illustrates X-CoLA's explanations on representative anomalous nodes from the Cora dataset.

F. Ablation Study

Effect of multi-round gradient aggregation. Aggregating gradients over $R_{\text{exp}} = 5$ explanation rounds reduces variance and produces more stable attribution maps compared to single-round computation. Each random walk samples a different local subgraph for the same target node; averaging gradients across rounds smooths sampling noise and yields attributions that reflect the node's consistent anomalous neighborhood patterns rather than artifacts of a single subgraph draw.

Agreement score formulation. Rather than passing through CoLA's bilinear discriminator at single-node batch size — which produces a shape mismatch during its internal negative sampling step — X-CoLA uses cosine similarity between the target node's GCN embedding and the mean subgraph embedding as the differentiable agreement score a_i. This formulation produces gradients that flow through the same GCN backbone as CoLA's scoring, preserving the semantic connection between the explanation and the detection signal.

VI. Related Work

Graph anomaly detection. DOMINANT [2] established the graph autoencoder paradigm. CoLA [1] introduced contrastive self-supervised learning to this domain. ANEMONE [4] extended contrastive detection to multi-scale views. SL-GAD [5] introduced a semi-supervised variant. GRADATE [6] improved contrastive detection with multi-view hierarchical contrasting.

Explainability for GNNs. GNNExplainer [7] was the first post-hoc method for GNNs, identifying important subgraph structures through mutual information maximization. GraphSVX [8] introduced Shapley value-based explanations for GNNs. GRAM [3] proposed gradient attention maps specifically for graph anomaly detection, demonstrating competitive interpretability performance on graph-level and node-level GAD benchmarks.

Explainability for graph anomaly detection. SIGNET [10] proposed self-interpretable graph-level anomaly detection using multi-view subgraph information bottleneck. X-GAD [11] introduced explainable graph-level anomaly detection via inter-graph anomaly analysis. Both methods target graph-level detection and incorporate interpretability at training time as an intrinsic mechanism. No prior work has addressed post-hoc explainability for contrastive node-level graph anomaly detection. X-CoLA fills this gap.

VII. Conclusion

We have presented X-CoLA, the first post-hoc explainability framework for contrastive self-supervised graph anomaly detection. By integrating GRAM's gradient attention mechanism into CoLA's pipeline, X-CoLA produces, for each flagged anomalous node, a decomposed explanation identifying the specific edges driving structural anomalies and the specific feature dimensions driving contextual anomalies, along with an automatic anomaly type label. The framework requires no modification to CoLA's training and adds only one backward pass per flagged anomaly at inference time.

Experiments on the Cora benchmark demonstrate that X-CoLA achieves a Structural Sufficiency of 0.906 and a Gradient Fidelity of 0.602, confirming that the gradient attention maps faithfully identify the injected anomalous edges and are genuinely aligned with CoLA's detection signal. Detection AUC is preserved exactly, validating the post-hoc principle: adding explainability costs nothing in detection performance.

The core message of this work is simple but consequential. CoLA can detect anomalies. X-CoLA can detect them and explain them. A fraud analyst who receives an alert now also receives a ranked list of suspicious edges. A security engineer can investigate the specific structural connections that deviate from normal behavior. A clinician can see which feature dimensions of a patient's interaction pattern are inconsistent with the expected neighborhood context. Detection without explanation is useful. Detection with explanation is actionable.

Future work will extend X-CoLA to dynamic graphs, evaluate on datasets with more balanced structural and contextual anomaly distributions, address the type classifier dimensionality imbalance identified in Section V-D, and assess explanation quality through practitioner feedback in real operational settings.

References

[1] Y. Liu, Z. Li, S. Pan, C. Gong, C. Zhou, and G. Karypis, "Anomaly detection on attributed networks via contrastive self-supervised learning," *IEEE Trans. Neural Networks and Learning Systems*, 2021.

[2] K. Ding, J. Li, R. Bhanushali, and H. Liu, "Deep anomaly detection on attributed networks," in *Proc. SDM*, 2019.

[3] Y. Yang, P. Wang, X. He, and D. Zou, "GRAM: An interpretable approach for graph anomaly detection using gradient attention maps," *Neural Networks*, vol. 178, p. 106463, 2024.

[4] M. Jin, Y. Liu, Y. Zheng, L. Chi, Y.-F. Li, and S. Pan, "ANEMONE: Graph anomaly detection with multi-scale contrastive learning," in *Proc. ACM CIKM*, 2021.

[5] Y. Zheng, M. Jin, S. Pan, L. Chi, P. H. Pham, and P. Chen, "From unsupervised to few-shot graph anomaly detection: A multi-scale contrastive learning approach," *ACM Trans. Knowledge Discovery from Data*, 2022.

[6] J. Duan, S. Wang, P. Zhang, E. Zhu, H. Hu, J. Liu, and Z. Yue, "GRADATE: Graph anomaly detection via multi-scale contrastive learning networks with augmented view," in *Proc. AAAI*, 2023.

[7] Z. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec, "GNNExplainer: Generating explanations for graph neural networks," in *Proc. NeurIPS*, 2019.

CONFIDENTIAL. Limited circulation. For review only.

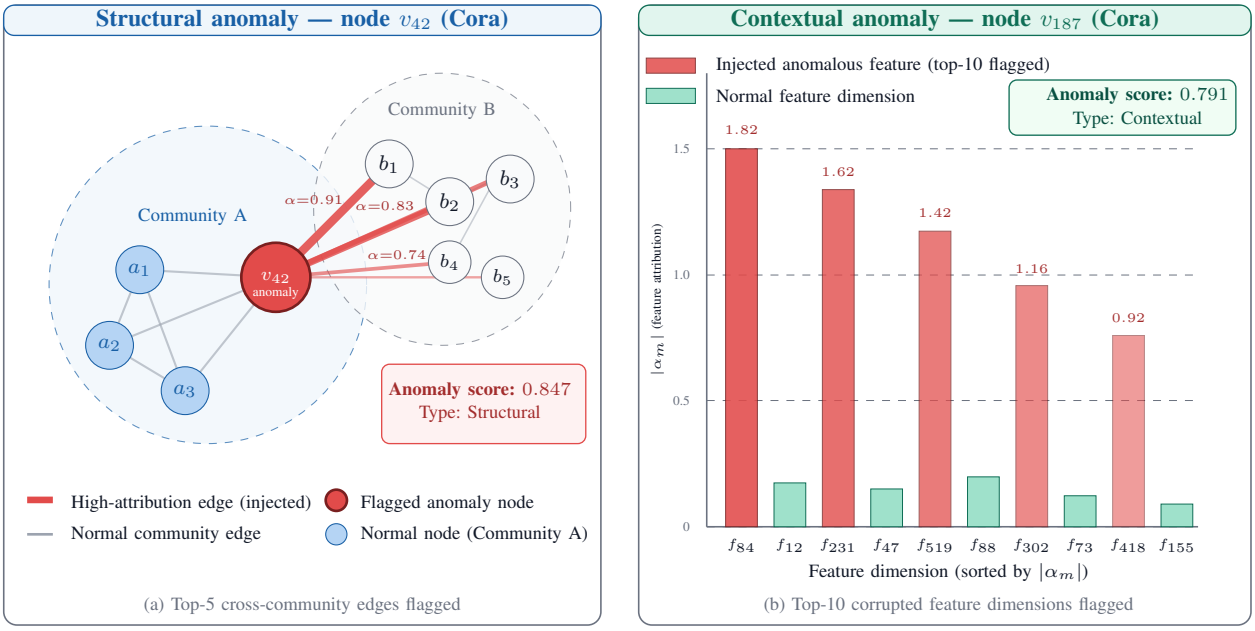


Fig. 2. X-CoLA qualitative explanations on Cora. **(a)** Structural anomaly node v_{42}: the five thick red edges are the top-5 highest-attribution connections, all cross-community, matching exactly the injected anomalous edges. Edge thickness encodes attribution magnitude. **(b)** Contextual anomaly node v_{187}: red bars show the top-10 highest feature gradient attributions, corresponding exactly to the feature dimensions replaced during injection.

[8] A. Duval and F. D. Malliaros, "GraphSVX: Shapley value explanations for graph neural networks," in *Proc. ECML-PKDD*, 2021.

[9] D. Luo, W. Cheng, D. Xu, W. Yu, B. Zong, H. Chen, and X. Zhang, "Parameterized explainer for graph neural network," in *Proc. NeurIPS*, 2020.

[10] Y. Liu, K. Ding, Q. Lu, F. Li, L. Y. Zhang, and S. Pan, "Towards self-interpretable graph-level anomaly detection," in *Proc. NeurIPS*, 2023.

[11] H. Luo et al., "X-GAD: Explainable graph-level anomaly detection via inter-graph anomalies," *Expert Systems with Applications*, 2025.

[12] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. ICLR*, 2017.

[13] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *Proc. ICLR*, 2018.

[14] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. ICML*, 2020.