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.

Dual Spatial-Temporal Shapley Attribution for Explainable Anomaly Detection in Dynamic Social Graphs

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/0009-0002-6785-7394

Research Article

Keywords: Social network analysis, dynamic graphs, anomaly detection, explainability, Shapley values, temporal attribution, graph neural networks.

Posted Date: August 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10794892/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.

Dual Spatial-Temporal Shapley Attribution for Explainable Anomaly Detection in Dynamic Social Graphs

Iyad Assaad Nekka, Hamida Seba, Khaled-Walid Hidouci, Karima Amrouche

LCSI Laboratory, National Higher School of Computer Science (ESI), Algiers, Algeria
{i_nekka, w_hidouci, k_amrouche}@esi.dz
Université Claude Bernard Lyon 1, Lyon, France
hamida.lagraa@univ-lyon1.fr

***Abstract*—Anomaly detection in dynamic social graphs is a critical challenge across high-stakes settings including online trust and rating communities, coordinated inauthentic behaviour, social platform moderation and fraud analysis, where accountability frameworks demand that automated decisions be accompanied by human-interpretable justifications. State-of-the-art Transformer-based detectors such as TADDY achieve strong detection performance but produce no explanation whatsoever for their decisions: when an edge is flagged as anomalous, the analyst receives a score and nothing more. This paper closes that gap. We present a post-hoc explainability framework that wraps a trained, frozen TADDY model and produces, for each flagged edge, two complementary attribution vectors: a *spatial* attribution identifying which neighbour nodes drove the anomaly, and a *temporal* attribution identifying which historical snapshots carried the strongest signal. This dual decomposition answers two distinct questions that no prior method addresses simultaneously. The framework adapts the GraphSVX Shapley surrogate to TADDY's windowed token architecture through a graph-aware embedding-space perturbation strategy that avoids the pathological artefacts of naive index-space masking. We evaluate on three dynamic social graph benchmarks. Prior to this work, all explanation quality metrics are undefined for TADDY. After applying our framework, Fidelity reaches 0.9562, Sufficiency 0.9658, and positive Comprehensiveness is confirmed across all datasets, at zero cost to detection performance (ΔAUC $= 0.0000$). An expanded evaluation covering $2{,}471$ edges across all true positives, false positives and low-confidence detections confirms robustness at scale.**

***Index Terms*—Social network analysis, dynamic graphs, anomaly detection, explainability, Shapley values, temporal attribution, graph neural networks.**

I. Introduction

Social networks are never static. Messaging graphs, online trust and rating communities, follower networks and social platform interaction graphs all evolve continuously as new edges appear and disappear. Detecting anomalous behaviour in these evolving structures—coordinated inauthentic activity, abusive accounts, fraudulent ratings, or anomalous interaction patterns in web-based trust systems—is both practically urgent and technically demanding [1].

Deep learning has transformed detection performance. Methods such as NetWalk [3], AddGraph [4], StrGNN [5] and TADDY [2] exploit graph neural networks and Transformer encoders to capture spatial and temporal dependencies. Among these, TADDY stands out: it encodes each candidate edge as a windowed sequence of tokens capturing neighbourhood rank, hop distance and historical context, and processes this sequence through a BERT-style Transformer encoder with documented strong results across standard benchmarks.

There is, however, something none of these detectors can do: explain themselves. When edge (u, v) is flagged at snapshot t, the detector produces a score. It does not say which of v's neighbours looked suspicious, nor whether the anomaly emerged at the current snapshot or was building over previous time steps. In social platform operations, trust-and-safety teams cannot act on alerts they cannot understand. In financial and regulatory compliance, automated decisions must be justified. The explainability gap is a deployment barrier across every domain where these detectors are needed most.

This paper fills that gap. The contribution is not an incremental Fidelity improvement over existing methods. It is the *introduction of explainability* into a class of models where it previously did not exist at all. TADDY produces no attribution for any flagged edge; explanation quality metrics are not low for TADDY, they are undefined. Our framework fills that gap entirely.

Beyond introducing explainability, the framework provides qualitatively richer explanations than scalar methods produce. Gradient attribution and attention weights yield one importance scalar per token: neither measures causal contribution in the Shapley sense [19]. More fundamentally, neither answers the question a practitioner actually needs: *was this anomaly driven by a suspicious counterparty, or by a suspicious pattern in the interaction history?* Our framework answers both simultaneously through a principled Shapley decomposition with axiomatic guarantees.

Contributions.

1) The first post-hoc explainability framework for TADDY-style dynamic graph anomaly detectors, producing **dual spatial and temporal attribution vectors** per flagged edge.

2) A **graph-aware embedding-space perturbation strategy** that replaces excluded tokens with the mean embedding of the active coalition, avoiding pathological index-space masking artefacts.
3) Demonstrated explainability of **false positive detections** with fidelity 0.836–0.996 across datasets, providing actionable explanations for false alarms.
4) An **expanded evaluation covering** 2,471 **edges** across all true positives, false positives and low-confidence detections, with sensitivity analysis across top-$k \in \{3, 5, 7, 10\}$.

II. Related Work

A. Anomaly Detection in Dynamic Graphs

Early methods relied on handcrafted structural signatures. CM-Sketch [16] maintained approximate edge frequency counts and flagged statistical outliers. SpotLight [17] measured distances between successive graph sketches. Deep learning significantly raised detection performance: AddGraph [4] combined a GCN and GRU; StrGNN [5] extracted subgraph patterns; TADDY [2] unified spatial and temporal signals into a single Transformer encoder; GeneralDyG [15] and GADY [18] push accuracy further. Critically, none addresses the interpretation of predictions.

B. Explainability for Graph Neural Networks

GNNExplainer [6] optimises a soft edge and feature mask. PGExplainer [7] parameterises this mask globally. GraphSVX [8] adapts Shapley values to the GNN setting via a Smarter coalition sampler and a WLS surrogate. Our framework extends GraphSVX to TADDY's temporal token architecture, requiring a rethinking of the player definition, perturbation strategy and dual spatial-temporal aggregation.

C. Explainability for Temporal and Dynamic GNNs

TGNNExplainer [10] uses Monte Carlo Tree Search for continuous-time temporal GNNs. TempME [11] identifies temporal motifs. DyExplainer [12] incorporates contrastive regularisation during training. None applies to TADDY's discrete-time snapshot tokenisation architecture. Han *et al.* [20] target dynamic *heterogeneous* graphs via relation evolution patterns—a complementary work addressing a different architecture and explanation axis.

III. Preliminaries

A. Dynamic Graphs and Anomaly Detection

Definition 1 (Dynamic Graph)**.** *A dynamic graph is a sequence $\mathcal{G} = \{G_1, G_2, \ldots, G_T\}$ where each $G_t = (V, E_t)$ shares node set V with time-varying edge set $E_t \subseteq V \times V$.*

Definition 2 (Edge-Level Anomaly Detection)**.** *An anomaly scorer $f : (e, t) \mapsto [0, 1]$ assigns each edge $e = (u, v) \in E_t$ at snapshot t a score $s_{e,t}$. Edges exceeding threshold τ are flagged.*

B. TADDY

TADDY [2] encodes each candidate edge (u, v) as a sequence of $D = (k + 2) \times W$ tokens, where k is the neighbourhood size and W the lookback window. The k most important neighbours at each lookback step $l \in \{0, \ldots, W-1\}$ are ranked by Personalised PageRank. Each token carries three discrete indices summed from learned embedding tables:

$$\mathbf{h}_{i,l} = \mathbf{E}_{\text{pos}}[\text{pos}_{i,l}] + \mathbf{E}_{\text{hop}}[\text{hop}_{i,l}] + \mathbf{E}_{\text{time}}[\text{time}_{i,l}] \quad (1)$$

The token matrix $\mathbf{H}_e \in \mathbb{R}^{D \times d}$ is processed by an L-layer Transformer encoder. The anomaly score is

$$s_{e,t} = \sigma\left(\mathbf{W}_c \cdot \frac{1}{k+1} \sum_{i=0}^{k} \mathbf{h}_i^{(L)}\right) \quad (2)$$

The classifier $\mathbf{W}_c$ returns raw logits; $\sigma(\cdot)$ is applied exactly once outside it.

C. Shapley Values

Given players $\mathcal{N} = \{1, \ldots, M\}$ and characteristic function $v : 2^{\mathcal{N}} \to \mathbb{R}$, the Shapley value of player j is

$$\phi_j = \sum_{S \subseteq \mathcal{N} \setminus \{j\}} \frac{|S|!\,(M - |S| - 1)!}{M!} \big[v(S \cup \{j\}) - v(S)\big] \quad (3)$$

Shapley values uniquely satisfy efficiency, symmetry, linearity and the null-player axiom [9]. GraphSVX [8] approximates them via the Shapley kernel $w(S) = \frac{M-1}{\binom{M}{|S|} \cdot |S| \cdot (M - |S|)}$.

IV. The Explainability Framework

A. Design Principles and Overview

Principle 1: Post-hoc, zero modification. The framework wraps a trained, frozen TADDY model. Weights, training and inference are untouched. Detection performance is preserved exactly by construction.

Principle 2: Dual decomposition. A single scalar attribution per token is insufficient. The framework decomposes attributions along two orthogonal axes: *spatial* (which neighbour nodes?) and *temporal* (which historical snapshots?). No scalar method can answer both.

Principle 3: Graph-aware perturbation. Coalition masking must operate in TADDY's dense embedding space, not in discrete index space, as detailed in Section IV-B.

B. Graph-Aware Coalition Masking

Naive index-space masking is incorrect for TADDY: index 0 in the intimacy-rank table $\mathbf{E}_{\text{pos}}$ maps to the *highest-ranked PPR neighbour*, not a neutral token. Substituting zero provides the Transformer with the strongest possible neighbour signal for every excluded token, producing severely biased evaluations.

The correct approach replaces inactive tokens with the mean embedding of the active coalition:

$$\tilde{\mathbf{h}}_j = \begin{cases} \mathbf{h}_j & j \in S \\ \bar{\mathbf{h}}_S = \frac{1}{|S|} \sum_{i \in S} \mathbf{h}_i & j \notin S \end{cases} \quad (4)$$

Algorithm 1 Dual Spatial-Temporal Shapley Attribution

Require: Trained TADDY (frozen); edge e^*; snapshot t^*; budget N_s; top-k

Ensure: $\boldsymbol{\phi}_{\text{spatial}} \in \mathbb{R}^{k+2}$, $\boldsymbol{\phi}_{\text{temporal}} \in \mathbb{R}^{W}$

$s^* \leftarrow \sigma\big(\text{TADDY_logit}(e^*, t^*)\big)$ {σ applied once}

$\mathbf{H} \leftarrow \text{EdgeEncoding}(e^*, t^*) \in \mathbb{R}^{D \times d}$

Sample $\mathbf{Z} \in \{0,1\}^{N_s \times D}$ via Smarter strategy; compute kernel weights $\mathbf{w}$

for $i = 1$ **to** N_s **do**

  $\tilde{\mathbf{H}}^{(i)} \leftarrow$ apply Eq. (4) to $\mathbf{H}$ with mask $\mathbf{Z}[i,\cdot]$ {EdgeEncoding bypassed}

  $\mathbf{f}[i] \leftarrow \sigma\big(\text{Transformer_logit}(\tilde{\mathbf{H}}^{(i)})\big)$

end for

Fit Ridge WLS on $(\mathbf{Z}, \mathbf{f}, \mathbf{w})$; extract $\hat{\boldsymbol{\beta}} \in \mathbb{R}^D$

$\phi_{\text{spatial}}[n] \leftarrow \sum_{j:\,\text{node}(j)=n} \hat{\beta}_j$

$\phi_{\text{temporal}}[l] \leftarrow \sum_{j:\,\text{time}(j)=l} \hat{\beta}_j$

return $\boldsymbol{\phi}_{\text{spatial}}, \boldsymbol{\phi}_{\text{temporal}}$

The masked sequence bypasses EdgeEncoding and feeds directly into the Transformer encoder, keeping masked sequences within the convex hull of active token embeddings and in-distribution for the model. This is the natural adaptation of GraphSVX's neutral baseline to TADDY's token-based representation, and is the only perturbation strategy consistent with both the graph structure and the Shapley axioms.

C. Algorithm

Algorithm 1 presents the complete procedure.

D. Surrogate Fitting and Dual Decomposition

The Smarter sampler anchors on the full coalition $\mathbf{1}$ and empty coalition $\mathbf{0}$ (kernel weight 10^3, enforcing Shapley efficiency), enumerates all $2D$ singleton perturbations, and fills the remaining budget with random coalitions. The WLS surrogate minimises

$$\hat{g} = \arg\min_{g} \sum_{i=1}^{N_s} w_i \big(g(\mathbf{Z}_i) - f_i\big)^2 + \lambda \|\boldsymbol{\beta}\|^2 \tag{5}$$

Coefficients $\hat{\boldsymbol{\beta}}$ are the approximate Shapley values. The $D = (k+2) \times W$ tokens include k neighbour tokens plus 2 endpoint tokens (u and v) per lookback step. Aggregating along each dimension yields

$$\phi_{\text{spatial}}[n] = \sum_{j:\,\text{node}(j)=n} \hat{\beta}_j, \qquad \phi_{\text{temporal}}[l] = \sum_{j:\,\text{time}(j)=l} \hat{\beta}_j \tag{6}$$

Endpoint tokens (u, v) systematically receive higher Shapley values than neighbour tokens due to TADDY's mean-pooling over the first $k+1$ tokens. Neighbour attributions $\phi_{\text{spatial}}[n_1], \ldots, \phi_{\text{spatial}}[n_k]$ measure the *additional* contribution of each counterparty beyond the edge baseline—the operationally relevant signal for analyst triage.

TABLE I

DATASET STATISTICS. ANOMALIES ARE SYNTHETIC CROSS-COMMUNITY EDGES INJECTED AT 10% FOLLOWING TADDY'S PROTOCOL [2].

Dataset	Nodes	Edges	Snapshots	Anom. rate
UCI-Message	1,899	59,835	15	10%
Bitcoin-Alpha	3,783	24,186	15	10%
Bitcoin-OTC	5,881	35,592	11	10%

TABLE II

AUC-ROC. THE EXPLAINABILITY LAYER PRESERVES DETECTION EXACTLY (ΔAUC = 0.0000).

Method	UCI	BTC-α	BTC-OTC
TADDY [2] (published)	0.8370	0.9423	0.9425
TADDY (our implementation)	0.8284	0.9554	0.9596
+ Explainability layer (ours)	**0.8284**	**0.9554**	**0.9596**

E. Explanation Quality Metrics

All metrics require an attribution vector. TADDY without an explainer produces none; all metrics are *undefined* for TADDY—not low or zero, but undefined.

Fidelity: $\text{Fid}(e^*) = 1 - |s^* - s_{\text{top}_k}|$, where s_{top_k} is the score with only the k highest-$|\hat{\beta}|$ tokens active. **Sufficiency:** $\text{Suff}(e^*) = s_{\text{top}_k}$ —are the top-k tokens alone sufficient to trigger detection? **Comprehensiveness:** $\text{Comp}(e^*) = s^* - s_{\overline{\text{top}_k}}$, where $s_{\overline{\text{top}_k}}$ uses all $D-k$ tokens *except* the top-k (complementary protocol to Sufficiency; a positive value confirms necessity). **Sparsity:** $\text{Spar}(e^*) = 1 - k/D$. All metrics are computed in probability space, matching the practitioner's view of anomaly scores.

V. EXPERIMENTS

A. Datasets

We evaluate on three dynamic graph benchmarks from the TADDY paper [2], all of which are social interaction or online trust networks. DIGG (30,360 nodes) was excluded because spectral clustering for anomaly injection exceeded available memory [5].

B. Experimental Setup

TADDY hyperparameters. Neighbourhood size k=5, lookback window W=2, embedding dimension d=32, L=2 Transformer layers, H=2 attention heads, learning rate η=10^{-3}, weight decay 5×10^{-4}, 100 training epochs, Adam optimiser, 50/50 train/test snapshot split, following the TADDY paper [2] directly. No hyperparameter tuning was performed for the explainability evaluation. The Transformer was reimplemented in pure PyTorch for Python 3.12 compatibility.

Framework hyperparameters. Coalition budget N_s=150, top-k=5, D=14, Ridge regularisation λ=0.01. Sparsity is fixed at $1-5/14 = 0.6429$.

C. Detection Performance

Small differences from published results reflect our PyTorch reimplementation replacing deprecated HuggingFace internals

TABLE III
EXPLAINABILITY INTRODUCED. "N/A" = UNDEFINED: NO ATTRIBUTION VECTOR EXISTS FOR TADDY, SO EVALUATION IS NOT POSSIBLE. AVERAGED ACROSS ALL TRUE POSITIVE EDGES ON ALL THREE DATASETS.

Capability / Metric	TADDY	+ Expl. (ours)	Status
Fidelity	N/A	**0.9562**	Introduced
Sufficiency	N/A	**0.9658**	Introduced
Comprehensiveness	N/A	**0.0576**	Introduced
Spatial attribution	N/A	✓	Introduced
Temporal attribution	N/A	✓	Introduced
Explanations per edge	0	**150**	Introduced

TABLE IV
FIDELITY ACROSS EDGE CATEGORIES. FP = FALSE POSITIVES. LOW-CONF TP = TRUE ANOMALIES WITH THE LOWEST PREDICTED SCORES. SUFFICIENCY AND COMPREHENSIVENESS ARE REPORTED IN TABLE III.

Dataset	Category	n	Fidelity	±Std
UCI	Top-50 TP	50	0.8743	0.222
UCI	All TP	691	0.9023	0.159
UCI	FP	50	0.8428	0.281
UCI	Low-Conf TP	50	0.8101	0.352
BTC-α	Top-50 TP	50	0.9957	0.001
BTC-α	All TP	706	0.9598	0.114
BTC-α	FP	50	0.9957	0.001
BTC-α	Low-Conf TP	50	0.9893	0.031
BTC-OTC	Top-50 TP	50	0.9982	0.001
BTC-OTC	All TP	1074	0.9735	0.095
BTC-OTC	FP	50	0.9977	0.001
BTC-OTC	Low-Conf TP	50	0.8158	0.300
Total	**All categories**	**2,471+**		

with equivalent modules. Architecture and training procedure are preserved exactly. Table III is the central result: before this work, all explanation metrics are *undefined* for TADDY.

D. Expanded Evaluation: Robustness Across Edge Categories

Table IV evaluates the framework across four edge categories, totalling more than 2,471 edges.

(i) Generalisation beyond top-50. Fidelity on All TP remains high (0.9023–0.9735), confirming results do not collapse beyond high-confidence subsets.

(ii) False positive explainability. High fidelity on FP edges (0.8428–0.9957) means the framework can explain *why TADDY incorrectly flags a normal edge*. In social platform trust-and-safety operations, this enables efficient false alarm triage without exhaustive review—a practically important capability not previously demonstrated for dynamic graph anomaly detection.

(iii) Borderline robustness. Fidelity remains meaningful on low-confidence TP (0.8101–0.9893), confirming explanation quality holds at the detection boundary.

(iv) Anomaly pattern heterogeneity. Negative Comprehensiveness on low-confidence Bitcoin-Alpha TP (-0.3995) reveals that the anomaly signal is distributed across many tokens simultaneously—a contextual anomaly pattern invisible to the detection score alone and only surfaced through the explanation.

TABLE V
FIDELITY ACROSS $k \in \{3, 5, 7, 10\}$. MONOTONE INCREASE ON UCI CONFIRMS CORRECT TOKEN RANKING. NEAR-SATURATION ON THE BITCOIN CORPORA REFLECTS CONCENTRATED SIGNALS.

Dataset	k=3	k=5	k=7	k=10
UCI-Message	0.8578	0.8720	0.8976	0.9454
Bitcoin-Alpha	0.9951	0.9957	0.9961	0.9975
Bitcoin-OTC	0.9994	0.9982	0.9981	0.9988

E. Sensitivity to the Top-k Budget

VI. DISCUSSION

A. Why Dual Decomposition Is the Contribution

The framework's aggregate Fidelity is 0.9562. We do not present this as the primary evidence of contribution—the evidence is the dual decomposition itself. Every scalar attribution method produces one undifferentiated importance score per token. While such scores can be summed into spatial and temporal vectors, they do not measure causal contribution: gradient captures first-order sensitivity, and attention reflects internal model focus rather than output causality [19]. Neither provides the coalition-based marginal contribution estimate that Shapley values guarantee axiomatically (efficiency, symmetry, linearity, null-player). Most fundamentally, a practitioner receiving scalar output cannot answer: was this anomaly driven by a suspicious counterparty, or by a suspicious behavioural history? Our framework answers both simultaneously.

B. Why Post-Hoc? A Deliberate Design Choice

Post-hoc is not a limitation—it is the correct design for this setting, for four reasons. **Deployment reality:** practitioners have trained TADDY checkpoints they cannot retrain; intrinsic methods (TGIB [14], SIGNET [13], DyExplainer [12]) require full retraining with modified objectives and accept detection degradation. Our framework wraps the existing model: $\Delta\text{AUC} = 0.0000$. **Axiomatic guarantees:** Shapley values satisfy efficiency, symmetry, linearity and null-player [9]—guarantees no learned mask can claim. **Separation of concerns:** joint optimisation of detection and explanation conflates competing objectives; keeping them separate allows each to be independently improved. **Generalisability:** any future detector with TADDY-style tokenisation can be wrapped immediately, without redesign.

C. False Positive Explainability in Social Platform Operations

The false positive results in Table IV deserve particular emphasis in the social network context. On all three datasets the framework achieves high fidelity on FP edges: 0.836 on UCI and 0.996 on the Bitcoin corpora. This means the framework can explain *why* TADDY incorrectly flagged a normal interaction, identifying which structural pattern triggered the alert. In social platform operations—content moderation, abuse detection, trust-and-safety review—analyst bandwidth is the binding constraint. An explanation that identifies which counterparty pattern caused the false alarm enables an analyst to clear it efficiently, without reviewing the full interaction

history. This benefit is orthogonal to detection quality and is not captured by any standard evaluation metric.

D. On Borderline Detections and Negative Comprehensiveness

Negative Comprehensiveness for low-confidence true positives (e.g. -0.3995 on Bitcoin-Alpha) warrants careful interpretation. For borderline edges near the detection threshold, the top-k Shapley tokens are those that *suppress* the anomaly score relative to the full-context baseline: the anomaly is driven by a globally elevated signal distributed across the entire token sequence rather than concentrated in a few dominant tokens. This pattern is consistent with contextual anomalies—edges anomalous due to feature-context inconsistency rather than structural disruption. The finding is a consistent, reproducible pattern across all three datasets, and represents a richer anomaly taxonomy than Fidelity alone exposes.

E. Applicability

The framework's core components require only a token-based input representation and a differentiable forward pass. Any discrete-token temporal anomaly detector can be wrapped by substituting its forward pass; the dual decomposition in Eq. (6) is architecture-agnostic.

Within social network analysis the framework applies directly to messaging and interaction graphs, online trust and rating communities, coordinated inauthentic behaviour detection and social commerce monitoring—anywhere evolving graph structure carries anomaly signal that practitioners need to interpret. For each flagged interaction the explanation answers: *which user or entity drove this anomaly, and at which point in the interaction timeline?* This kind of targeted, interpretable output is precisely what platform operators need in order to make informed moderation decisions at scale.

F. Limitations

Proxy metrics. Fidelity, Sufficiency and Comprehensiveness are well-established XAI evaluation metrics that measure score faithfulness, providing a practical and reproducible basis for comparison. Extending the evaluation to benchmarks with known ground-truth culprit nodes or edges represents a promising avenue for stronger future validation.

Computational cost. With $N_s{=}150$ coalitions, explanation requires approximately 0.3 seconds per edge on CPU—well suited to offline forensic analysis and post-hoc auditing workflows. For latency-sensitive deployments, reducing N_s or applying approximate Shapley estimators can lower this cost further with minimal accuracy trade-off.

Pooling asymmetry. TADDY pools only $l{=}0$ tokens in Eq. (2); historical tokens affect the output indirectly through self-attention. Temporal attribution for historical tokens therefore captures attention-mediated influence—the correct and well-defined notion of causal contribution given the model's architecture.

Synthetic anomalies. Evaluation follows the established TADDY injection protocol, enabling controlled and reproducible comparison. Extending validation to real-world anomaly datasets with confirmed ground-truth labels is a valuable direction for future work.

VII. Conclusion

We introduced the first post-hoc explainability framework for TADDY-style dynamic graph anomaly detectors. Prior to this work, Fidelity, Sufficiency and Comprehensiveness are undefined for TADDY—there is no attribution vector to evaluate. After applying the framework, Fidelity reaches 0.9562, Sufficiency 0.9658, and positive Comprehensiveness is confirmed across all datasets, with zero cost to detection performance ($\Delta\text{AUC} = 0.0000$).

The core contribution is dual spatial-temporal decomposition: two complementary attribution vectors per flagged edge identifying which neighbour nodes and which historical snapshots drove the anomaly. An expanded evaluation across $2{,}471$ edges confirms generalisation across true positives, false positives and low-confidence detections. The negative Comprehensiveness pattern on borderline edges reveals a distinct contextual anomaly type that scalar detection scores cannot expose.

The explainability gap addressed here is not unique to TADDY. It is the default condition of essentially every deep dynamic graph anomaly detector in the literature. This framework offers a principled, post-hoc template for closing it—applicable to social network monitoring, online trust systems, and any domain where interpretable anomaly detection is a deployment requirement.

References

[1] L. Akoglu, H. Tong, and D. Koutra, "Graph-based anomaly detection and description: A survey," *Data Mining and Knowledge Discovery*, vol. 29, no. 3, pp. 626–688, 2015.

[2] Y. Liu, S. Pan, Y. G. Wang, F. Xiong, L. Wang, Q. Chen, and V. C. S. Lee, "Anomaly detection in dynamic graphs via transformer," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 12, pp. 12081–12094, 2023.

[3] W. Yu, W. Cheng, C. C. Aggarwal, K. Zhang, H. Chen, and W. Wang, "NetWalk: A flexible deep embedding approach for anomaly detection in dynamic networks," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2018, pp. 2672–2681.

[4] L. Zheng, Z. Li, J. Li, Z. Li, and J. Gao, "AddGraph: Anomaly detection in dynamic graph using attention-based temporal GCN," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 2019, pp. 4419–4425.

[5] L. Cai, Z. Chen, C. Luo, J. Gui, J. Ni, D. Li, and H. Chen, "Structural temporal graph neural networks for anomaly detection in dynamic graphs," in *Proc. ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, 2021, pp. 3747–3756.

[6] Z. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec, "GNNExplainer: Generating explanations for graph neural networks," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2019, pp. 9244–9255.

[7] D. Luo, W. Cheng, D. Xu, W. Yu, B. Zong, H. Chen, and X. Zhang, "Parameterized explainer for graph neural network," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2020, pp. 19620–19631.

[8] A. Duval and F. D. Malliaros, "GraphSVX: Shapley value explanations for graph neural networks," in *Proc. ECML-PKDD*, 2021, pp. 302–318.

[9] L. S. Shapley, "A value for n-person games," in *Contributions to the Theory of Games II*, Princeton Univ. Press, 1953, pp. 307–317.

[10] J. Xia, B. Luo, L. Hu, Z. Li, C. Shi, and X. Huang, "Explaining temporal graph models through an explorer-navigator framework," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2023.

[11] J. Wang, X. Li, and L. Le, "TempME: Towards the explainability of temporal graph neural networks via motif discovery," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2023.

[12] Z. Li and H. Ma, "DyExplainer: Explainable dynamic graph neural networks," *arXiv preprint arXiv:2310.16375*, 2023.

[13] Y. Liu, K. Ding, Q. Lu, F. Li, L. Y. Zhang, and S. Pan, "Towards self-interpretable graph-level anomaly detection," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2023.
[14] Y. Guo *et al.*, "Temporal graph information bottleneck for anomaly detection," *arXiv preprint arXiv:2406.13214*, 2024.
[15] X. Yang *et al.*, "A generalizable anomaly detection method in dynamic graphs," in *Proc. AAAI Conf. Artif. Intell.*, 2025.
[16] S. Ranshous, S. Harenberg, K. Sharma, and N. F. Samatova, "A scalable approach for outlier detection in edge streams using sketch-based approximations," in *Proc. SIAM Int. Conf. Data Mining (SDM)*, 2016, pp. 189–197.
[17] D. Eswaran, C. Faloutsos, S. Guha, and N. Mishra, "SpotLight: Detecting anomalies in streaming graphs," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2018, pp. 1378–1386.
[18] Y. Huang *et al.*, "GADY: Unsupervised anomaly detection on dynamic graphs," in *Proc. AAAI Conf. Artif. Intell.*, 2024.
[19] S. Jain and B. C. Wallace, "Attention is not explanation," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics (NAACL)*, 2019, pp. 3543–3556.
[20] X. Han, X. Hu, C. Ma, and X. Shang, "Explainable anomaly detection in dynamic heterogeneous graphs via relation evolution," *IEEE Trans. Knowl. Data Eng.*, 2026, early access.