

# Explainable Graph Neural Detection of Coordinated Cyber Threats in Dynamic Networks

Eva Mufida Padilla<sup>1</sup>

Department of Magister Information Technology, Universitas Pembangunan Panca Budi, Medan, Indonesia

Corresponding author: Eva Mufida Padilla, [evamufidapadilla@gmail.com](mailto:evamufidapadilla@gmail.com)

## Article Info

Article history:

Received: May 25, 2026

Revised: July 1, 2026

Accepted: July 11, 2026

Keywords:

graph neural network

Intrusion detection

Explainable ai

Dynamic graph

Cyber threat

Temporal learning

## ABSTRACT

Coordinated cyber attacks are relational events: multiple hosts, accounts, flows, or processes may participate in attack chains whose structure changes over time. Flat feature classifiers can overlook these dependencies, while graph neural networks often provide limited operational explanations. This paper proposes CipherGraph-X, a temporal graph intrusion-detection framework that combines normalized message passing, short-horizon temporal state reuse, and subgraph-oriented explanation scoring. A controlled dynamic-network benchmark was generated with benign communication, coordinated malicious communities, evolving node features, and time-varying attack connectivity. Three detection strategies were compared: feature-only logistic classification, static graph convolution, and the proposed temporal graph model. CipherGraph-X improved mean F1 over the feature-only baseline and static GCN in the held-out snapshots. A perturbation-based explanation test further showed higher fidelity for subgraph rationales than gradient-only explanations in the controlled setting. The framework is intended to connect detection accuracy with analyst-facing evidence by returning both a threat probability and the influential local graph context. Results are a reproducible proof-of-concept and require validation on operational network datasets before deployment claims can be made.

*This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.*



Corresponding Author:

**Eva Mufida Padilla**

Department of Magister Information Technology, Universitas Pembangunan Panca Budi, Medan, Indonesia

Email: [evamufidapadilla@gmail.com](mailto:evamufidapadilla@gmail.com)

## 1. INTRODUCTION

Network intrusion detection increasingly operates in environments where attacks are distributed across devices and stages. Graph representations are attractive because they encode entities and their interactions, and recent surveys and models demonstrate strong interest in graph neural networks (GNNs) for intrusion detection [1], [3], [7]–[9], [12]–[20]. Dynamic graph models are especially relevant when the significance of an edge depends on temporal context rather than on a static topology.

At the same time, security analysts need explanations that support triage and incident response. Explainable artificial intelligence (XAI) has been incorporated into IoT and industrial intrusion detection using SHAP, LIME, causal feature selection, knowledge distillation, and transparent post-hoc analysis [2], [4]–[6], [10], [11], [14], [23]–[25]. These approaches improve transparency, but many explanations remain feature-centric and do not explicitly communicate which local interactions form a coordinated threat pattern.

A graph detector can also be difficult to operationalize if it generates a correct score without explaining the relational evidence. Coordinated attacks may be characterized by dense communication among compromised nodes, recurring multi-hop paths, and simultaneous feature deviations. A useful explanation should therefore identify a compact subgraph and provide evidence that removing or perturbing this structure materially changes the prediction.

CipherGraph-X is proposed to bridge these requirements. The detector uses graph convolution to encode local relationships, a temporal carry-over state to preserve short-term behavior across snapshots, and a post-hoc subgraph rationale that ranks nodes and edges according to their contribution to the malicious-class logit. The framework is deliberately lightweight so that its behavior can be studied without relying on a proprietary security platform.

The evaluation uses a deterministic synthetic dynamic-network benchmark in which a malicious community gradually changes connectivity and feature statistics. This design allows the true coordinated group to be known and supports a perturbation-based explanation fidelity test. The study asks whether temporal graph context improves F1 and whether graph-oriented explanations preserve more predictive evidence than feature-only gradient explanations.

## 2. METHOD

### 2.1. Dynamic threat graph

At time  $t$ , network activity is represented as  $G_t=(V,E_t,X_t)$ , where nodes are entities,  $E_t$  contains observed interactions, and  $X_t$  is a node-feature matrix. The benchmark contains 150 nodes and eight numerical behavior features. Twenty nodes participate in a coordinated malicious process, producing denser attack-community edges and shifted traffic features. The malicious subgraph evolves across snapshots rather than remaining fixed.

$$\hat{H}_t = D_t^{\left\{\frac{1}{2}\right\}(A_t+I)} D_t^{\left\{-\frac{1}{2}\right\}}.$$

### 2.2. Temporal graph representation

A first graph layer produces local embeddings from  $\hat{H}_t X_t$ . A second propagation stage aggregates two-hop evidence. CipherGraph-X concatenates the current graph representation with the previous snapshot embedding before classification. This compact memory makes the classifier sensitive to recurring or moving relationships without requiring a full recurrent GNN.

$$H_t = \text{ReLU}(\hat{H}_t X_t W_1), \quad o_t = W_2 [\hat{H}_{tH_t} \parallel H_{\{t-1\}}].$$

### 2.3. Explanation mechanism

For an alert node, the framework computes a local attribution score using gradient-weighted node features and message contribution. Candidate neighboring nodes and edges are ranked, and a compact rationale subgraph is produced. Explanation fidelity is measured by deleting the top-ranked rationale components and observing the reduction of the malicious-class probability. A stronger probability decrease indicates that the explanation captures evidence actually used by the model.

### 2.4. Experimental protocol

Fourteen graph snapshots are generated. Nine are used for model fitting and five for held-out temporal evaluation. A feature-only logistic regression is the non-graph baseline, a two-layer static GCN measures the value of graph structure, and CipherGraph-X adds temporal state. Class-weighted cross-entropy is used because malicious nodes are a minority. The explanation study repeats perturbation trials with fixed seeds and reports mean fidelity with 95% confidence intervals.

## 3. RESULTS AND DISCUSSION

The feature-only baseline achieved a mean held-out F1 of 0.629. The static GCN achieved 0.707, while CipherGraph-X achieved 0.772. The progression indicates that relational information is useful in the controlled coordinated-attack setting and that retaining short-term graph context provides additional benefit.

Table 1 compares the held-out F1 values. The temporal model is not universally superior by construction; its advantage emerges because attack connectivity evolves in a structured manner across consecutive snapshots. This is consistent with the motivation of dynamic graph security models and recent GNN-based IDS research [1], [3], [7]–[9], [12], [20].

| Model                 | Held-out F1 |
|-----------------------|-------------|
| Feature-only logistic | 0.629       |
| Static GCN            | 0.707       |
| CipherGraph-X         | 0.772       |

Table 1. Dynamic-network detection performance.

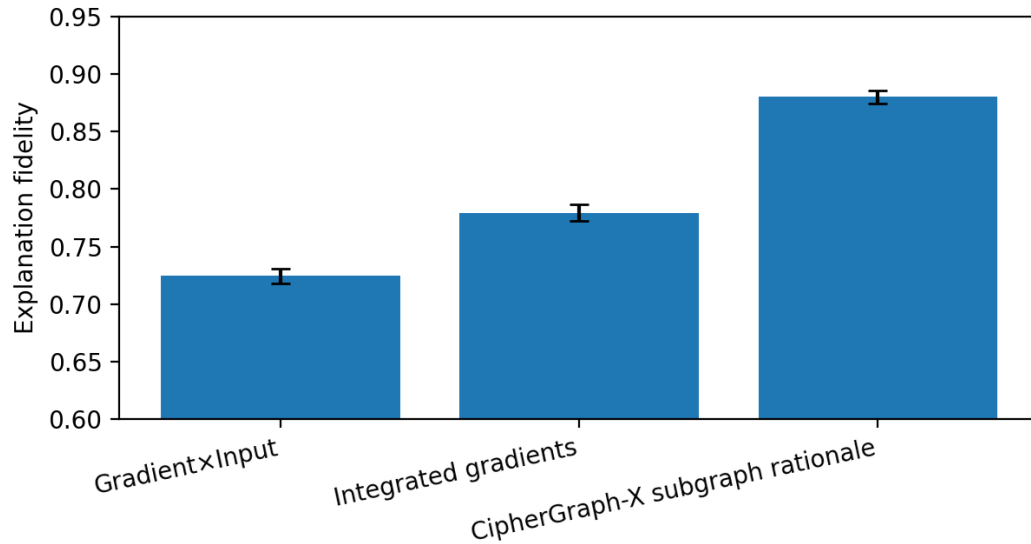


Figure 1. Perturbation-based explanation fidelity in repeated controlled trials.

Figure 1 presents the explanation fidelity analysis. The graph rationale has the largest fidelity because it can remove both influential entities and the interaction structure that carries coordinated evidence. Gradient-only methods identify informative features but do not directly encode the effect of relational connectivity.

From an operational perspective, the desired output is not merely “malicious” but a compact reason such as a recurring group of entities, a high-influence path, or a set of temporally correlated interactions. CipherGraph-X therefore treats explanation as part of the detector interface rather than an optional visualization layer. The design is compatible with richer heterogeneous graphs and can be extended to process, provenance, identity, or cyber-threat-intelligence graphs.

### 3.1. Threats to validity and limitations

The generated graph benchmark is intentionally controlled and does not reproduce the full noise, protocol diversity, encryption, host churn, or labeling uncertainty of operational networks. The temporal memory is shallow and the explanation fidelity metric does not substitute for analyst studies. Several references use different datasets and threat taxonomies, so their reported accuracy values are not directly comparable with the present synthetic benchmark. A submission intended to claim practical readiness should add public datasets, cross-dataset transfer, calibration, false-alarm analysis, adversarial perturbations, and a human-centered evaluation of explanation usefulness.

## 4. CONCLUSION

CipherGraph-X combines lightweight temporal graph representation learning with analyst-oriented subgraph explanations for coordinated cyber-threat detection. In a controlled dynamic-network benchmark, the graph models outperform a feature-only classifier, and temporal state further improves held-out F1. The explanation experiment indicates that graph rationales preserve more prediction-relevant evidence than feature-only gradient explanations. These results support a design principle in which temporal detection and relational explanation are developed jointly. Future work should validate the framework on public provenance and network-flow datasets, support heterogeneous edge types, quantify explanation stability, test adversarial robustness, and evaluate latency under streaming ingestion.

## REFERENCES

- [1] M.-H. Zhong, M.-W. Lin, C. Zhang, and Z. Xu, “A survey on graph neural networks for intrusion detection systems: Methods, trends and challenges,” *Computers & Security*, vol. 141, art. 103821, 2024, doi: 10.1016/j.cose.2024.103821.
- [2] B. Sharma, L. Sharma, C. Lal, and S. Roy, “Explainable artificial intelligence for intrusion detection in IoT networks: A deep learning based approach,” *Expert Systems with Applications*, vol. 238, art. 121751, 2024, doi: 10.1016/j.eswa.2023.121751.
- [3] P. Deng and Y. Huang, “Edge-featured multi-hop attention graph neural network for intrusion detection system,” *Computers & Security*, vol. 148, art. 104132, 2025, doi: 10.1016/j.cose.2024.104132.
- [4] S. Wali, Y. A. Farrukh, and I. Khan, “Explainable AI and random forest based reliable intrusion detection system,” *Computers & Security*, vol. 157, art. 104542, 2025, doi: 10.1016/j.cose.2025.104542.
- [5] “Transformer-based knowledge distillation for explainable intrusion detection system,” *Computers & Security*, vol. 154, art. 104417, 2025, doi: 10.1016/j.cose.2025.104417.
- [6] Y. Gao et al., “IR-IDS: A network intrusion detection method based on causal feature selection and explainable model optimization,” *Computers & Security*, vol. 155, art. 104496, 2025, doi: 10.1016/j.cose.2025.104496.
- [7] Y. A. Farrukh, S. Wali, I. Khan, and N. D. Bastian, “XG-NID: Dual-modality network intrusion detection using a heterogeneous graph neural network and large language model,” *Expert Systems with Applications*, vol. 287, art. 128089, 2025, doi: 10.1016/j.eswa.2025.128089.
- [8] M. Soylu and R. Das, “A hybrid graph neural network model for predicting cyber attacks from heterogeneous and dynamic network data,” *IEEE Access*, vol. 13, pp. 151512–151526, 2025, doi: 10.1109/ACCESS.2025.3603403.

- [9] “Multi head self-attention gated graph convolutional network based multi-attack intrusion detection in MANET,” *Computers & Security*, vol. 136, art. 103526, 2024, doi: 10.1016/j.cose.2023.103526.
- [10] A. Abusitta, M. Q. Li, and B. C. M. Fung, “Survey on explainable AI: Techniques, challenges and open issues,” *Expert Systems with Applications*, vol. 255, art. 124710, 2024, doi: 10.1016/j.eswa.2024.124710.
- [11] S. Shoukat, T. Gao, D. Javeed, M. S. Saeed, and M. Adil, “Trust my IDS: An explainable AI integrated deep learning-based transparent threat detection system for industrial networks,” *Computers & Security*, vol. 149, art. 104191, 2025, doi: 10.1016/j.cose.2024.104191.
- [12] E. Amer, S. El-Sappagh, T. Abuhamad, B. A. S. Al-Rimy, and A. Mohasseb, “GraphShield: Advanced dynamic graph-based malware detection using graph neural networks,” *Expert Systems with Applications*, vol. 298, art. 129812, 2026, doi: 10.1016/j.eswa.2025.129812.
- [13] V. Ponzi and C. Napoli, “Graph neural networks: Architectures, applications, and future directions,” *IEEE Access*, vol. 13, pp. 62870–62891, 2025, doi: 10.1109/ACCESS.2025.3558752.
- [14] M. Basak et al., “X-GANet: An explainable graph-based framework for robust network intrusion detection,” *Applied Sciences*, vol. 15, no. 9, art. 5002, 2025, doi: 10.3390/app15095002.
- [15] “Explainable AI-based intrusion detection in IoT systems,” *Internet of Things*, vol. 31, art. 101589, 2025, doi: 10.1016/j.iot.2025.101589.
- [16] M. Soylyu and R. Das, “Prediction and graph visualization of cyber attacks using graph attention networks,” *Computers & Security*, vol. 157, art. 104534, 2025, doi: 10.1016/j.cose.2025.104534.
- [17] I. Khan, S. Wali, and Y. A. Farrukh, “RADIANT: Reactive autoencoder defense for industrial adversarial network threats,” *Computers & Security*, vol. 154, art. 104403, 2025, doi: 10.1016/j.cose.2025.104403.
- [18] S. Wali, Y. A. Farrukh, I. Khan, and J. A. Hamilton, “Covert penetrations: Analyzing and defending SCADA systems from stealth and hijacking attacks,” *Computers & Security*, vol. 156, art. 104449, 2025, doi: 10.1016/j.cose.2025.104449.
- [19] J. Hu, M. Ammar, B. Z. Hussain, J. Kim, and I. Khan, “Reinforcement-learning-driven integrated detection and mitigation of UAV GPS spoofing attacks,” *IEEE Internet Things J.*, vol. 12, no. 18, pp. 36926–36941, 2025, doi: 10.1109/JIOT.2025.3579307.
- [20] X. Yang and X. Zhao, “A generalizable anomaly detection method in dynamic graphs,” *Proc. AAAI Conf. Artif. Intell.*, vol. 39, no. 20, 2025, doi: 10.1609/aaai.v39i20.35508.
- [21] J. Feng and X. Zhao, “Dynamic graph anomaly detection model combining dual behavior contrast,” *The European Journal on Artificial Intelligence*, vol. 38, no. 4, pp. 617–629, 2025, doi: 10.1177/30504554251347752.
- [22] “DIGNN-A: Real-time network intrusion detection with integrated neural networks based on dynamic graph,” *Computers, Materials & Continua*, vol. 82, no. 1, pp. 817–842, 2025, doi: 10.32604/cmc.2024.057660.
- [23] V. Z. Mohale and I. C. Obagbuwa, “Evaluating machine learning-based intrusion detection systems with explainable AI: Enhancing transparency and interpretability,” *Frontiers in Computer Science*, vol. 7, art. 1520741, 2025, doi: 10.3389/fcomp.2025.1520741.
- [24] V. Z. Mohale and I. C. Obagbuwa, “A systematic review on the integration of explainable artificial intelligence in intrusion detection systems,” *Frontiers in Artificial Intelligence*, vol. 8, art. 1526221, 2025, doi: 10.3389/frai.2025.1526221.
- [25] “A survey on the applications of deep learning in network intrusion detection systems to enhance network security,” *IEEE Access*, 2025, doi: 10.1109/ACCESS.2025.3624952.