International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

An Explainable Graph Neural Network Framework for Detecting Multi-Stage Cyberattacks in Enterprise Networks

Dr. Satya Venkata Surya Narayana Murthy
Mphasis Corporation
41 Madison Ave 35th floor, New York, NY 10010
nimusha73@gmail.com

Abstract

Multi-stage intrusions leave weak, scattered traces: a suspicious process on one workstation, an unusual authentication on another, a privilege change on a server. Analysts must connect these events into a coherent attack chain without mistaking legitimate administrative workflows for an attacker. This paper presents Sentinel-XAI, a research design that represents enterprise activity as a typed temporal graph of users, hosts, processes, services and network sessions, learns from it with relational attention-based message passing under strict temporal cutoffs, and tests every reconstructed path with deletion-based explanation fidelity before it reaches an analyst. The time-bounded attack chain is the unit of analysis, which makes the decision boundary and accountability explicit. We formalize graph construction, message passing, chain scoring and explanation fidelity; define evaluation with chain recall, event-level precision, detection delay and deletion fidelity; and specify comparisons against sequence classifiers, static graph embeddings and rule-based correlation. A worked scenario follows a phishing-originated execution through credential reuse to lateral movement. The design is restricted to technologies available in 2020. No experimental results are claimed; the contribution is an implementable, independently testable research design and an explicit account of its validity threats, including the risk of reading attention weights as causal explanations.

Keywords: graph neural networks, temporal graphs, explainable AI, intrusion detection, attack-chain reconstruction, provenance analysis, MITRE ATT&CK

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

1. Introduction

Modern intrusions unfold in stages: initial access, execution, credential access, lateral movement and privilege escalation, as catalogued in the MITRE ATT&CK knowledge base [4]. Each stage may produce only a weak signal on a single host. Security operations centers therefore face a reconstruction problem: which of thousands of individually unremarkable events belong to the same campaign?

Provenance-based systems have shown that following causal dependencies between system events can recover attack stories [10], [12], [13] and reduce alert fatigue [11]. At the same time, graph neural networks (GNNs) learn from relational structure at scale [1], [2], [6]. Combining the two is attractive but raises a governance problem: a black-box score attached to an attack chain is of little use if analysts cannot verify which events drove it, and attention weights alone are not reliable explanations [16], [17].

This paper proposes Sentinel-XAI, whose contributions are:

1) A formulation in which the **time-bounded attack chain** is the unit of analysis, rather than the individual alert.
2) An architecture combining typed temporal graph construction, relational attention-based message passing with strict temporal cutoffs, and path-level explanation testing.
3) A mathematical specification of message passing, chain scoring, cost-sensitive decisions and deletion-based explanation fidelity.
4) A reproducible evaluation protocol with baselines, temporal partitions and threat analysis, presented without unperformed experiments.

Section II reviews related work, Section III states the research questions, Section IV the data model, Section V the architecture and Section VI the formalization. Sections VII–IX cover the illustrative scenario, implementation and evaluation protocol; Section X discusses validity and ethics, and Section XI concludes.

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

2. Background and Related Work

A. Provenance and Attack-Story Reconstruction

King and Chen introduced backtracking from a detection point through operating-system dependencies [10]. SLEUTH [13] and HOLMES [12] reconstruct attack scenarios in real time from audit data and map them to adversary stages, and NoDoze [11] ranks alerts by the anomaly of their provenance context. These systems rely largely on expert-defined rules and propagation policies.

B. Graph Representation Learning

Graph convolutional networks [7], inductive GraphSAGE [1] and graph attention networks [2] learn node representations from neighborhoods; relational GCNs extend this to typed edges [6], and temporal graph networks handle evolving graphs [9]. Static embeddings such as node2vec [23] ignore time. PyTorch Geometric made these models practical [8].

C. Sequence Models for Security Logs

Recurrent models such as LSTMs [24] underpin log anomaly detectors like DeepLog [22], which model each host's event sequence but not cross-host relationships.

D. Explanation and Evaluation

GNNExplainer identifies the subgraph most responsible for a prediction [3]. Whether attention weights constitute explanations is contested [16], [17], so Sentinel-XAI evaluates explanations by deletion. Evaluation practice in security also matters: temporal and spatial bias can inflate reported results [20], and the gap between laboratory and operational detection is well documented [18].

Research gap. Provenance systems explain but rely on hand-written rules; GNNs learn but are rarely tested for explanation fidelity at the level of a full attack chain. Sentinel-XAI is designed to close this gap.

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

3. Problem Formulation and Research Questions

The setting is an enterprise network with user identities, workstations, processes, services and network sessions. The objective is to identify attack chains early and produce incident narratives that analysts can defend. Four research questions are posed:

1) **RQ1 (effectiveness):** Does typed temporal message passing with path-level explanation improve chain identification compared with sequence classifiers, static graph embeddings and rule-based correlation?
2) **RQ2 (generalization):** How does performance change when the mix of events shifts, for example after a software rollout, and do temporal cutoffs and preserved provenance reduce the degradation?
3) **RQ3 (efficiency and governance):** Can chains be reconstructed within acceptable delay while keeping benign administrative workflows from being attributed to attackers?
4) **RQ4 (explainability):** Do the edges an explanation highlights actually drive the chain score, as measured by deletion fidelity?

Each question is a testable comparative proposition. The scope excludes universal intrusion detection and automated containment.

4. Data Model and Provenance

Events are converted into a typed temporal graph (Table I). Every edge carries a timestamp, a source-system identifier and the raw event reference, so each graph element can be traced to its origin.

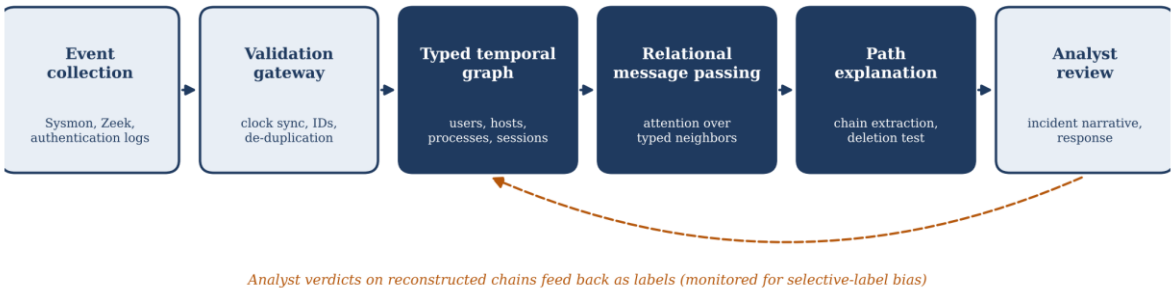
TABLE I. Graph schema: edge types, endpoints and sources

Edge type	Connects	Typical source (2020)
Process launch	User/process → process on a host	Sysmon process-creation events [21]

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

Authentication	User → host or service	Windows and domain authentication logs [14]
Lateral connection	Host → host via a network session	Zeek connection and protocol logs [15]
Privilege change	User or process → privileged group or token	Directory and endpoint audit events

Temporal integrity. A prediction made at time t_0 may use only edges with timestamps up to t_0, which prevents future events from leaking into the score.



Normalization. Host and user identifiers are reconciled across sources, clocks are synchronized, and missing log sources are recorded explicitly instead of producing silent gaps in the graph.

Governance. Only authorized telemetry is ingested; user identities can be pseudonymized for model development, and every transformation is logged.

Partitioning. Training, validation and test periods are separated in time, and attack campaigns are never split across partitions, following the temporal-consistency principles of [20]. The LANL multi-source dataset [14] is a suitable public starting point.

5. The Sentinel-XAI Architecture

Fig. 1 shows the pipeline. The three central stages carry the learning and explanation.

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

Fig. 1. Sentinel-XAI pipeline. Graph construction, message passing and path explanation (dark) form the core; analyst verdicts return as labels.

Local telemetry and schema mapping convert each participant's events into the shared taxonomy and validate timestamps, formats, authorization, and duplicates.

- **Event collection and validation** gather endpoint, network and authentication events and check clock synchronization, identifiers, authorization, and duplicates.
- **Typed temporal graph** construction builds nodes and timestamped, typed edges within a sliding window.
- **Relational message passing** computes node and edge representations with attention over typed neighbors, respecting the temporal cutoff.
- **Path explanation** extracts the highest-scoring time-respecting path and tests it by deletion before it is shown.
- **Analyst review** receives the chain, its evidence, and an incident narrative. Model confidence alone never authorizes containment.

Failure handling. If a log source is missing for a host, the chain is marked incomplete and scored conservatively; such fallbacks are monitored by source and network segment.

6. Mathematical Formulation

A. Typed Temporal Graph

Let G = (V, E, R, T) with nodes V, edges E, edge types R and timestamps T. For a decision time t_0, only $E_{t_0} = \{e \in E : t_e \leq t_0\}$ is visible.

B. Relational Message Passing

For node v at layer k, with typed neighborhoods $N_r(v)$ restricted to E_{t_0}, the update is [2], [6]:

$$h_v^{(k+1)} = \sigma\left(W_0 h_v^{(k)} + \sum_r \sum_{u \in N_r(v)} \alpha_{uvr} W_r h_u^{(k)} \right) \quad (1)$$

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

where the attention coefficient is computed as in graph attention networks [2]:

$$\alpha_{uvr} = \mathrm{softmax}_u\left(\mathrm{LeakyReLU}\left(a_r^T [W_r h_v \,\|\, W_r h_u]\right)\right) \quad (2)$$

C. Chain Scoring and Decision

A candidate chain $P = (e_1, \ldots, e_L)$ is a time-respecting path ($t_{e_1} \le \ldots \le t_{e_L}$). Its score aggregates edge scores $s(e)$ from the final layer:

$$S(P) = \frac{1}{L} \sum_{\ell=1..L} s(e_\ell) \quad (3)$$

The chain is escalated if $S(P) \ge \tau$, with τ chosen on validation data to minimize $C_{FN} \cdot FN(\tau) + C_{FP} \cdot FP(\tau)$ subject to analyst capacity K.

D. Explanation Fidelity and Evaluation Metrics

For an explanation $E_P \subset P$, deletion fidelity measures how much the score drops when the highlighted edges are removed [3]:

$$Fid(P) = S(P) - S(P \setminus E_P) \quad (4)$$

Chain recall is the fraction of ground-truth chains with at least a fraction ρ of their edges recovered, and detection delay is the time between the first malicious event and escalation:

$$R_{chain} = \left|\{C \in C_{true} : overlap(C, \hat{P}) \ge \rho\}\right| / |C_{true}|, \quad Delay = t_{alert} - t_{first} \quad (5)$$

These definitions do not establish that attention reflects causation. Graph incompleteness, attention mistaken for explanation and alert cascades can each invalidate an apparently convincing result.

7. Illustrative Scenario

Consider a user who opens a phishing attachment, after which harvested credentials are reused to reach a file server.

Step 1: **Define the observation boundary:** endpoint, authentication and network logs for the affected subnet are in scope; personal devices are not.

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

Step 2: **Build the graph:** Office application → script interpreter (process launch), workstation → file server (lateral connection), user account → server (authentication with a credential not normally used there).

Step 3: **Evaluate simple baselines:** a per-host sequence model flags the script launch but cannot link it to the server; a correlation rule fires only if both events match a pre-written signature.

Step 4: **Message passing and path extraction:** the time-respecting path from the attachment to the server receives a high score because each edge is unusual in context.

Step 5: **Explanation test:** deleting the credential-reuse edge sharply lowers the score, confirming that the explanation reflects what the model relied on; the chain is escalated with this evidence.

A legitimate administrator using remote management would follow a similar path but with routine edge context, illustrating why benign workflows must be represented in training data. This scenario is illustrative; no detection rate or runtime is inferred from it.

8. Implementation Considerations

Table II lists a period-appropriate technology stack (2020). These are examples, not a claim that an experiment was executed with them.

TABLE II. Candidate implementation stack (2020)

Component	Technology and role
Endpoint telemetry	Microsoft Sysmon for process, network, and registry events [21]
Network telemetry	Zeek (formerly Bro) connection and protocol logs [15]
Graph learning	PyTorch Geometric (2019) [8] with GraphSAGE [1] and GAT [2] layers
Training	Python 3.8, Adam optimizer [5]
Explanation	GNNExplainer-style masks [3] plus deletion testing

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

Engineering practice. A versioned data contract defines node and edge types. Parsing, graph construction, training, inference, explanation, and audit are separate modules [19]. Configuration management records versions, seeds, and thresholds; access control restricts who can view raw events or approve response.

Trade-offs. Longer graph windows capture slower campaigns but increase memory and delay; neighbor sampling [1] reduces cost but may omit rare edges.

9. Evaluation Protocol

Sentinel-XAI is compared with three baselines (Table III) using identical data and decision horizons.

TABLE III. Baselines and the question each comparison answers

Baseline	Question answered by the comparison
Sequence classifier (e.g., DeepLog [22])	Does cross-host relational structure add value beyond per-host sequences?
Static graph embedding (node2vec [23])	Does modeling time and edge types matter?
Rule-based correlation	Does learning improve on expert-written correlation rules?

Metrics. Chain recall, event-level precision, detection delay and deletion fidelity are reported with bootstrap confidence intervals. Accuracy is not reported because malicious events are rare.

Robustness and ablation. Tests remove log sources, shift windows, inject label noise and add benign administrative activity. Ablations remove edge types, temporal cutoffs and attention in turn.

Reporting standard. No dataset results are presented here. A valid evaluation should release code, partitions and exclusions, and report unfavorable outcomes.

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

10. Threats to Validity and Ethical Considerations

Internal validity. Missing log sources produce incomplete graphs; alert cascades can make one malicious event inflate scores of many neighbors.

Explanation validity. Attention weights are not explanations by default [16], [17]; only deletion-tested explanations are presented to analysts.

External validity. Red-team or synthetic attacks may differ from real campaigns, and enterprise environments drift [18], [20].

Feedback. Analyst verdicts on escalated chains are a biased sample of all activity; a random sample of unescalated chains is periodically reviewed.

Ethics. Activity graphs reveal detailed employee behavior. Access is restricted, identities are pseudonymized where possible, and monitoring follows organizational policy and law.

11. Conclusion and Future Work

Sentinel-XAI is a technically specified design for reconstructing multi-stage intrusion paths from heterogeneous enterprise events. It links the problem of connecting weak cross-host signals without blaming benign workflows to typed temporal graphs, relational message passing and tested path explanations.

The phishing-to-lateral-movement scenario shows how a deletion-tested chain can give analysts a defensible narrative. The contribution is an architecture and analytical argument, not a claim of demonstrated superiority. Future work will evaluate the protocol on public and authorized enterprise data, compare explanation methods and study long-term behavior under drift.

Acknowledgment

Technology and product names are used for identification and historical context only; no endorsement by their respective owners is implied.

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

References

1. W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Advances in Neural Information Processing Systems 30 (NIPS)*, Long Beach, CA, USA, 2017, pp. 1024–1034.
2. P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *Proc. 6th Int. Conf. Learning Representations (ICLR)*, Vancouver, Canada, 2018.
3. R. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec, "GNNExplainer: Generating explanations for graph neural networks," in *Advances in Neural Information Processing Systems 32 (NeurIPS)*, Vancouver, Canada, 2019, pp. 9240–9251.
4. B. E. Strom, A. Applebaum, D. P. Miller, K. C. Nickels, A. G. Pennington, and C. B. Thomas, "MITRE ATT&CK: Design and philosophy," MITRE Corporation, Tech. Rep. MTR180108, 2018.
5. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learning Representations (ICLR)*, San Diego, CA, USA, 2015.
6. M. Schlichtkrull, T. N. Kipf, P. Bloem, R. van den Berg, I. Titov, and M. Welling, "Modeling relational data with graph convolutional networks," in *Proc. 15th Extended Semantic Web Conf. (ESWC)*, Heraklion, Greece, 2018, pp. 593–607.
7. T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. 5th Int. Conf. Learning Representations (ICLR)*, Toulon, France, 2017.
8. M. Fey and J. E. Lenssen, "Fast graph representation learning with PyTorch Geometric," in *ICLR Workshop on Representation Learning on Graphs and Manifolds*, New Orleans, LA, USA, 2019.
9. E. Rossi, B. Chamberlain, F. Frasca, D. Eynard, F. Monti, and M. Bronstein, "Temporal graph networks for deep learning on dynamic graphs," arXiv:2006.10637, 2020.

10. S. T. King and P. M. Chen, "Backtracking intrusions," in *Proc. 19th ACM Symp. Operating Systems Principles (SOSP)*, Bolton Landing, NY, USA, 2003, pp. 223–236.
11. W. U. Hassan *et al.*, "NoDoze: Combatting threat alert fatigue with automated provenance triage," in *Proc. Network and Distributed System Security Symp. (NDSS)*, San Diego, CA, USA, 2019.
12. S. M. Milajerdi, R. Gjomemo, B. Eshete, R. Sekar, and V. N. Venkatakrishnan, "HOLMES: Real-time APT detection through correlation of suspicious information flows," in *Proc. IEEE Symp. Security and Privacy*, San Francisco, CA, USA, 2019, pp. 1137–1152.
13. M. N. Hossain *et al.*, "SLEUTH: Real-time attack scenario reconstruction from COTS audit data," in *Proc. 26th USENIX Security Symp.*, Vancouver, Canada, 2017, pp. 487–504.
14. A. D. Kent, "Comprehensive, multi-source cyber-security events," Los Alamos National Laboratory, Data set, 2015, doi: 10.17021/1179829.
15. V. Paxson, "Bro: A system for detecting network intruders in real-time," *Computer Networks*, vol. 31, no. 23–24, pp. 2435–2463, 1999.
16. S. Jain and B. C. Wallace, "Attention is not explanation," in *Proc. 2019 Conf. North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, Minneapolis, MN, USA, 2019, pp. 3543–3556.
17. S. Wiegreffe and Y. Pinter, "Attention is not not explanation," in *Proc. 2019 Conf. Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, 2019, pp. 11–20.
18. R. Sommer and V. Paxson, "Outside the closed world: On using machine learning for network intrusion detection," in *Proc. IEEE Symp. Security and Privacy*, Oakland, CA, USA, 2010, pp. 305–316.
19. D. Sculley *et al.*, "Hidden technical debt in machine learning systems," in *Advances in Neural Information Processing Systems 28 (NIPS)*, Montréal, Canada, 2015, pp. 2503–2511.

20. F. Pendlebury, F. Pierazzi, R. Jordaney, J. Kinder, and L. Cavallaro, "TESSERACT: Eliminating experimental bias in malware classification across space and time," in *Proc. 28th USENIX Security Symp.*, Santa Clara, CA, USA, 2019, pp. 729–746.
21. M. Russinovich and T. Garnier, "Sysmon," Microsoft Sysinternals, Software documentation, 2020.
22. M. Du, F. Li, G. Zheng, and V. Srikumar, "DeepLog: Anomaly detection and diagnosis from system logs through deep learning," in *Proc. ACM SIGSAC Conf. Computer and Communications Security (CCS)*, Dallas, TX, USA, 2017, pp. 1285–1298.
23. A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 855–864.
24. S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.