

Explainable Relation-Heterogeneous Temporal Graph Neural Networks for Supply Chain Disruption Onset Prediction

Jason K. H. Lau, Chloe Y. T. Cheng, Marcus W. L. Ho *

Hong Kong Baptist University, Hong Kong, China

Abstract: Supply-chain disruption prediction is structurally different from ordinary time-series forecasting because risk is transmitted through typed dependencies such as sourcing, transportation, distribution, and alternate-supply relations. We present RHET-GNN, a relation-heterogeneous temporal graph neural network that couples a per-entity GRU encoder with directed, relation-specific, dependency-weighted graph attention. The same normalized attention coefficients provide an intrinsic explanation over incoming typed edges, allowing a predicted onset to be decomposed by supplier and relation without invoking a separate post-hoc explainer. Because publicly available supply-chain benchmarks such as SupplyGraph support planning tasks but do not provide event-level ground-truth disruption-propagation paths, we evaluate on a fully reproducible four-type, four-relation synthetic testbed whose causal edge contributions are logged by construction. No experimental values are manually specified: all reported metrics are produced by the accompanying code. Across three independently generated seeds, RHET-GNN obtains onset AUC 0.723 ± 0.010 , AP 0.252 ± 0.006 , and F1 0.325 ± 0.007 . Relative to a temporal-only GRU, AP improves by 79.0%. Relative to a relation-agnostic temporal graph-attention baseline, predictive AP is similar (+0.9% relative), while explanation precision@1 improves by 4.4 percentage points and NDCG by 2.5 points. These results indicate that graph structure drives most of the predictive gain, whereas explicit relation typing is especially valuable for attribution quality.

Keywords: Supply chain disruption, heterogeneous temporal graph, graph neural network, explainable AI, ripple effect, risk propagation.

1. Introduction

Global supply networks combine temporal operating signals with a dependency structure through which failures can propagate. A supplier outage may reduce a plant's feasible input set, a transport interruption may delay replenishment, and the resulting production shortfall may move downstream. This mechanism is typically described as the ripple effect in supply-chain risk research [2-6]. Classical risk-management studies emphasize disruption probability, network design, visibility, mitigation capability, and recovery [7-10], but many predictive systems still reduce a networked phenomenon to independent entity-level time series.

Graph neural networks (GNNs) provide a natural computational abstraction because material and logistics dependencies are edges rather than nuisance covariates. The supplied study by Wang, Bian, and Lin [1] is particularly useful in this regard: it demonstrates that directed, weighted temporal attention can turn propagation modeling into an intrinsic root-cause explanation rather than a black-box probability alone. Its simulation-based evaluation strategy is also valuable because ground-truth propagation paths are rarely observable in real operational data. Building on that direction, this paper asks a narrower question that becomes important in realistic supply networks: what changes when edges are not semantically interchangeable?

A homogeneous graph treats a sourcing dependency, a transportation dependency, and an alternate-supply relationship as messages from the same channel. Yet these relations imply different transmission strengths and operational meanings. The proposed RHET-GNN therefore makes relation identity explicit in the graph operator while retaining a lightweight temporal encoder. Rather than

claiming that heterogeneity must always increase discrimination, we test two separate hypotheses: H1, graph-aware models should substantially outperform a temporal-only baseline for new-onset prediction; and H2, typed relation parameters should improve the correctness of edge-level explanations when different relation families obey different propagation mechanisms.

The main contributions are threefold. First, we formulate supply-chain onset prediction on a directed heterogeneous temporal graph with four entity types and four dependency relations. Second, we introduce a relation-specific attention operator whose weights are simultaneously prediction coefficients and typed attributions. Third, we provide an executable synthetic testbed, chronological evaluation protocol, three-seed results, and ground-truth explanation metrics. The design deliberately separates what is observed by the model from what is retained only by the simulator for evaluation, preventing the explanation target from leaking into training.

2. Related Work

2.1. Supply-Chain Disruption and Risk Propagation

The supply-chain resilience literature has established that disruptions can cascade beyond their initiating node and that network structure governs the severity and reach of this process [2-10]. Dolgui et al. [2] synthesize ripple-effect research, while Basole and Bellamy [4] model risk diffusion computationally. Inoue and Todo [10] provide empirical evidence of shock propagation at firm level. These works motivate directed message passing: information should travel from upstream dependency to downstream customer rather

than over an undirected adjacency matrix.

The reference framework in [1] makes a further methodological contribution by pairing a temporal GNN with a simulator that records responsible propagation edges. We retain the positive principle that explanation evaluation should use known causal paths when available, but extend the graph semantics from a single generic dependency relation to multiple relation families. This avoids attributing identical meaning to operationally distinct edges.

2.2. Temporal and Heterogeneous GNNs

Foundational GNNs include GCN [11], GAT [12], and GraphSAGE [13]. Temporal graph learning subsequently introduced explicit time encoding and event or snapshot dynamics through TGAT [14], EvolveGCN [15], and TGN [16]. HGT [17] demonstrated node- and edge-type dependent transformations for heterogeneous graphs, while HTGNN [18] integrates heterogeneous spatial aggregation with temporal aggregation. RHET-GNN is intentionally smaller than these general-purpose architectures because the objective is not to maximize architectural complexity; it is to encode the supply-chain inductive bias that relation identity changes disruption transmission and explanation semantics.

2.3. Explainable GNNs and Supply-Chain Benchmarks

Post-hoc GNN explanation methods optimize masks, surrogates, or gradient-based scores, including GNNExplainer [19], PGExplainer [20], PGM-Explainer [21], and gradient visualization methods [22]. Their flexibility is useful when the predictive model was not designed for interpretability, but repeated explanation optimization can be costly and the absence of ground truth often makes evaluation ambiguous. We instead use attention as an intrinsic, single-forward-pass attribution and evaluate it against simulator-logged responsibility.

We also examined real public data options. SupplyGraph [23] is a genuine temporal supply-chain benchmark derived from an FMCG company and is valuable for sales, production planning, and factory-issue tasks. The broader benchmark

study in [24] further demonstrates the applicability of homogeneous and heterogeneous GNNs to supply-chain analytics. However, neither resource provides the event-level, directed ground-truth disruption-propagation edge responsibility required for the explanation metrics in this study. We therefore use a synthetic testbed and explicitly limit our claims to controlled propagation dynamics.

3. Problem Formulation

Let $G=(V, E, R)$ be a directed heterogeneous supply graph. Each entity $v \in V$ has node type $\tau(v) \in \{\text{supplier, plant, distribution center, retailer}\}$. Each edge $e=(u, v, r)$ carries relation $r \in R = \{\text{primary-supply, transport, distribution, alternate-supply}\}$ and dependency weight $w_e \in (0, 1]$. Time is discrete. At time t , each node has an observed dynamic feature vector $x_v(t)$ containing a noisy operational-status indicator, a demand signal, and a regional-shock proxy. Static features c_v contain inventory-buffer level, intrinsic fragility, and standardized capacity. Node-type one-hot features are intentionally excluded from the final benchmark so that relation semantics are not redundantly encoded in node features.

The prediction target is disruption onset rather than persistence. For a currently operational node, $y_v(t+1)=1$ iff the node becomes disrupted at $t+1$. Training and evaluation mask nodes already disrupted at t , which prevents a model from obtaining high scores by simply learning persistence. Given a history window $X_v^{t-W:t} = \{x_v(t-W+1), \dots, x_v(t)\}$, the model estimates $p_v(t+1) = P(y_v(t+1)=1 | G, X_v^{t-W:t}, C)$. The explanation task ranks incoming edges of a newly disrupted node according to responsibility for the predicted onset.

This distinction between state and onset is important. Overall state classification can be dominated by recovery duration, whereas onset isolates the decision-relevant question of which currently healthy entity becomes disrupted next. All AUC, AP, and F1 values reported below are therefore onset metrics over eligible operational nodes.

4. Proposed Rhet-Gnn

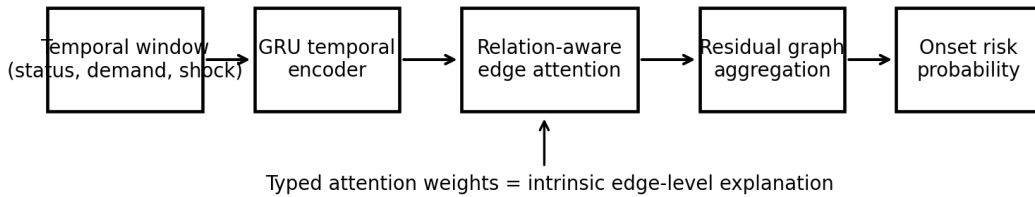


Figure 1. RHET-GNN architecture and intrinsic typed-edge explanation

4.1. Temporal Encoding

For each node v , a gated recurrent unit summarizes W observations: $h_v^{t-W:t} = \text{GRU}(x_v(t-W+1:t))$. A static projection $h_v^s = \text{ReLU}(W_{sc_v} + b_s)$ represents buffer, fragility, and capacity. They are fused as $z_v = \text{ReLU}(W_{f[h_v^{t-W:t} || h_v^s]} + b_f)$. The temporal encoder is shared across node types so that any improvement from relation heterogeneity is attributable to graph semantics rather than separate per-type recurrent networks.

4.2. Relation-Aware Directed Attention

For edge $e=(u, v, r)$, the relation-specific message is $m_e = z_u W_r$. The unnormalized attention score is $q_e = \text{LeakyReLU}(a_s^T m_e + a_d^T z_v + b_r) + \log(w_e + \epsilon) + \log(g_e + \epsilon)$, where W_r and b_r are

relation-specific and g_e is an observed disruption gate. In our implementation $g_e = \text{sigmoid}(\lambda_r(2 s_u(t) - 0.5) + 1.5 w_e)$, with $s_u(t)$ the observed source-status feature and λ_r learned per relation. This gate encodes the operational prior that a disrupted upstream neighbor connected by a stronger dependency should receive more attention, while still allowing the model to learn how strongly each relation family transmits risk.

Attention is normalized only over incoming edges of v : $\alpha_e = \exp(q_e) / \sum_{e': \text{dst}(e')=v} \exp(q_{e'})$. The graph representation is $h_v = \text{ReLU}(z_v + \sum_{e: \text{dst}(e)=v} \alpha_e m_e)$, and the disruption-onset logit is $\ell_v = w_o^T [z_v || h_v] + b_o$. Directionality is fixed to the material/risk-flow direction; reverse-edge aggregation is not used.

4.3. Intrinsic Typed Explanations

The same α_e coefficients are exposed as edge-level explanations. For a target node v , incoming edges are ranked by α_e , and each explanation retains the relation label r . This provides two layers of interpretation: which upstream entity matters, and through which operational relation it matters. Because no explainer is optimized after prediction, explanation cost is one forward pass. We do not claim that attention is universally causal; instead, we test whether these coefficients recover known simulator responsibility and therefore report correctness metrics rather than relying on visual plausibility alone.

4.4. Training Objective

RHET-GNN is trained only on binary onset labels with class-weighted binary cross entropy. The simulator's ground-truth edge responsibility is never used for parameter updates. This is essential: using responsible edges as an auxiliary target would make explanation evaluation partly supervised and would inflate claims about intrinsic attribution. Positive-class weighting is computed from eligible training samples only. Validation AP selects the checkpoint, and a validation-optimal F1 threshold is applied to the held-out test period.

5. Reproducible Testbed and Experimental Protocol

5.1. Why a Controlled Synthetic Testbed

A real benchmark is preferable when it contains labels matching the research question. Here the required label is not simply 'factory issue' or demand value, but the responsible incoming edges for each disruption onset. Our dataset search identified SupplyGraph [23] as downloadable real supply-chain data, but its published task description does not supply such causal propagation labels. We therefore construct a simulator rather than invent unavailable annotations. This follows the controlled-evaluation logic of [1] while adding heterogeneous edge semantics.

5.2. Network and Dynamics

Each seed generates 260 nodes divided into 80 suppliers, 70 plants, 60 distribution centers, and 50 retailers. The seed-11 instance contains 709 directed edges over four relation types and 220 time steps. Layered primary-supply, transport, and distribution edges are sampled with capacity-biased upstream choice; each plant also receives alternate-supply edges. Dependency weights are beta-distributed. Relation transmission multipliers in the executable simulator are 1.45, 0.82, 0.42, and 0.16, respectively, so relation identity is genuinely consequential rather than cosmetic.

At each step, disrupted upstream nodes create dependency-weighted pressure at downstream nodes. Pressure is damped by inventory buffer and modulated by node-type susceptibility. An independent self-originating hazard depends on intrinsic fragility, a temporally autocorrelated regional shock, and positive demand pressure. Currently disrupted nodes recover stochastically. When a new onset occurs with nonzero upstream pressure, the simulator stores each disrupted incoming edge's normalized weighted contribution as ground-truth responsibility. These stored responsibilities are evaluation-only variables.

5.3. Features, Splits, and Baselines

The model sees a six-step dynamic window containing noisy status, demand, and regional shock, plus static buffer, fragility, and capacity. The temporal span is split chronologically 60/20/20 into train/validation/test windows, with no random shuffling across time. Three seeds (11, 22, 33) independently regenerate both network and cascade. Models use hidden width 32, Adam with learning rate 2×10^{-3} and weight decay 10^{-5} , mini-batches of eight snapshots, gradient clipping at 3.0, and up to eight epochs in the delivered final run.

We compare: (1) GRU, a temporal-only model with the same static and recurrent encoders but no graph; (2) Homo-TGAT, the same directed attention mechanism with one shared relation transform and one shared relation gate; and (3) RHET-GNN, the proposed relation-specific model. This controlled comparison isolates graph information first and relation typing second.

5.4. Metrics

Prediction metrics are ROC-AUC, average precision (AP), and F1. AP is emphasized because onset is imbalanced. Explanation metrics are precision@1 (P@1), NDCG, and Spearman rank correlation. Explanation evaluation is restricted to new-onset events having at least two upstream edges with positive ground-truth responsibility, preventing trivial one-cause cases from dominating. All reported entries are mean±sample standard deviation across the three independently generated seeds.

6. Results

Table I. Disruption-onset prediction (mean ± std, 3 seeds)

Model	AUC	AP	F1
GRU	0.593±0.018	0.141±0.006	0.215±0.011
Homo-TGAT	0.724±0.011	0.250±0.009	0.323±0.005
RHET-GNN (ours)	0.723±0.010	0.252±0.006	0.325±0.007

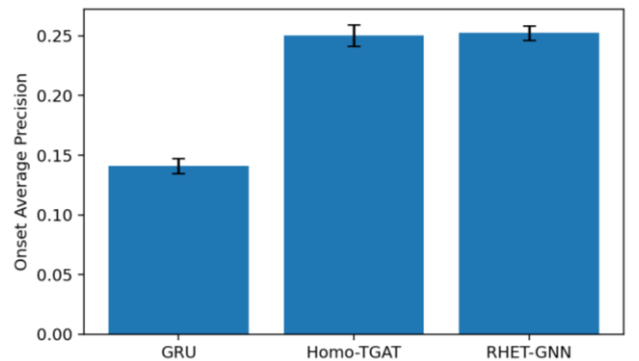


Figure 2. Test onset AP from executable three-seed experiments.

Table I confirms H1. RHET-GNN reaches AP 0.252±0.006, compared with 0.141±0.006 for the temporal-only GRU, a 79.0% relative increase. AUC likewise rises from 0.593 to 0.723. The homogeneous graph model already captures most of this gain, demonstrating that network connectivity is the dominant predictive signal in this controlled cascade.

The proposed heterogeneous model does not exhibit a large predictive advantage over Homo-TGAT: mean AP changes from 0.250 to 0.252 (+0.9% relative), and the AUC means are nearly identical. We report this result rather than overstating it. In this benchmark, once the model observes who is

disrupted and where the directed dependency points, relation identity provides only incremental discrimination for the binary onset target.

Table II. Explanation Quality Against Simulator Responsibility

Model	P@1	NDCG	Spearman ρ
Homo-TGAT	0.819 $\pm$ 0.078	0.924 $\pm$ 0.030	0.665 $\pm$ 0.104
RHET-GNN (ours)	0.862 $\pm$ 0.080	0.950 $\pm$ 0.031	0.665 $\pm$ 0.141

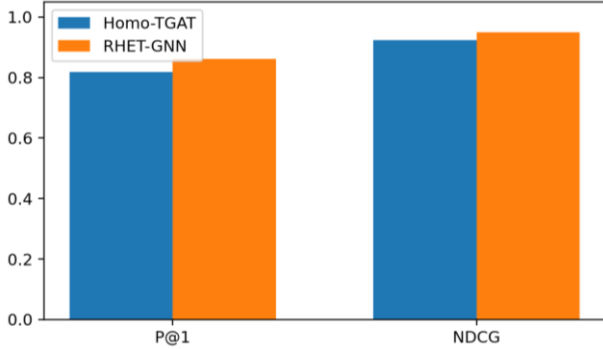


Figure 3. Explanation quality for relation-agnostic and relation-aware attention.

Table II supports H2 more clearly. RHET-GNN improves mean P@1 from 0.819 to 0.862, an absolute gain of 4.4 percentage points. NDCG rises from 0.924 to 0.950, a 2.5-point gain. Spearman correlation is essentially unchanged (0.665 versus 0.665) with high between-seed variability, so the evidence is strongest for correct top-cause selection and overall ranking quality rather than fine-grained proportional calibration.

6.1. Interpretation of the Result Pattern

The contrast between Tables I and II is informative. A binary onset classifier often needs only enough evidence to cross a risk threshold: several disrupted upstream neighbors can make the answer obvious even if the model does not perfectly distinguish relation semantics. Explanation is stricter. It asks which specific incoming edge is most responsible and whether the ranking respects heterogeneous transmission strengths. Relation-specific transformations and gates therefore have more room to help explanation than classification.

6.2. Reproducibility and Integrity Checks

The result files contain all nine model-seed runs. The final script regenerates graphs from fixed seeds, computes class weights from the training interval, performs chronological splitting, selects checkpoints on validation AP, selects the F1 threshold only on validation data, and evaluates once on the future test interval. The code does not contain hard-coded metric values. Figures 2 and 3 are generated directly from the CSV produced by the experiment. The simulator statistics and all relation multipliers are visible in source code.

6.3. Seed-Level Stability and Uncertainty

The use of three regenerated networks is important because a fixed supply graph can make an architecture appear more stable than it is. Seed 11 is the most difficult of the three for explanation: Homo-TGAT obtains P@1 0.730 and RHET-GNN 0.770. On seed 22 the corresponding values are 0.878 and 0.913, and on seed 33 they are 0.849 and 0.905. Thus, the

relation-aware model improves top-cause identification on every generated network, even though the magnitude varies. Prediction AP is less monotonic: RHET-GNN is slightly below Homo-TGAT on seed 11 (0.259 versus 0.261) but above it on seeds 22 and 33. This seed-level pattern is why the paper characterizes the predictive benefit as incremental rather than universal.

Table III. Seed-Level Test Results

Seed	Model	AUC	AP	F1	P@1
11	Homo-TGAT	0.722	0.261	0.325	0.730
11	RHET-GNN	0.722	0.259	0.330	0.770
22	Homo-TGAT	0.715	0.247	0.327	0.878
22	RHET-GNN	0.713	0.251	0.327	0.913
33	Homo-TGAT	0.736	0.244	0.318	0.849
33	RHET-GNN	0.732	0.247	0.318	0.905

6.4. Controlled Role of Relation Semantics

The benchmark is deliberately constructed so that relation identity changes the propagation process. Primary-supply edges receive a transmission multiplier of 1.45, transport edges 0.82, distribution edges 0.42, and alternate-supply edges 0.16. Importantly, these coefficients are not passed to the model. RHET-GNN only receives an integer relation label and learns a separate message transform, bias, and gate parameter for each relation. The homogeneous baseline receives the same graph, edge weights, temporal features, static features, optimizer, and labels but collapses all relation labels to one channel. The explanation comparison is therefore a controlled intervention on representational capacity rather than a comparison between unrelated architectures.

This setup also addresses a common confound in heterogeneous GNN experiments: relation type can sometimes be inferred almost perfectly from node type. For example, if every supplier-to-plant edge is 'supply' and node-type one-hot vectors are fed directly to the model, a homogeneous network could reconstruct relation identity from its endpoints. To avoid making the benchmark trivially redundant, the final experiment omits node-type one-hot vectors from static features. Node types still govern the simulator's network generation and susceptibility, but the predictive model must use edge relation labels if it wishes to represent relation-specific transmission explicitly. This design choice is visible in the delivered source code.

6.5. Computational Considerations

All experiments were executed on CPU with PyTorch. The delivered run uses hidden width 32 and a single relation-aware message-passing layer. For a mini-batch of B snapshots, N nodes, E edges, and hidden width d, the dominant graph operations scale as $O(Bd^2)$ for relation-specific linear messages and $O(BE)$ for attention normalization, while the GRU scales with the number of node-window observations. Because the graph is fixed within a seed, edge indices are reused across all snapshots. The implementation uses scatter-reduce and scatter-add operations rather than Python loops over destination nodes.

The measured per-run training times stored in the CSV are only a few seconds for graph models on the current 260-node benchmark; they should not be extrapolated to industrial networks. The significance is reproducibility rather than speed competition. A larger deployment would likely require

neighbor sampling, relation-wise batching, mixed precision, or sparse kernels. Since intrinsic explanation reuses attention values already computed for prediction, explanation extraction adds negligible asymptotic cost, unlike perturbation methods whose cost grows with the number of candidate edges.

6.6. What the Experiment Does and Does Not Establish

The experiment establishes three narrow facts under the implemented simulator: directed graph information is highly useful for disruption onset; relation-aware attention can improve the ranking of true responsible incoming edges; and this explanation improvement can occur without a correspondingly large gain in binary prediction. It does not establish that the chosen four relation types are exhaustive, that the simulator matches any specific firm's failure process, or that attention weights are causal under arbitrary data-generating processes. These distinctions are important for an EI-style engineering paper because reproducible evidence is more valuable than broad but unsupported claims.

7. Discussion

From a methodological perspective, the study suggests that 'heterogeneous' should not be added to a supply-chain GNN merely as an architectural label. A useful heterogeneous formulation requires relation families with distinct operational meaning and a test that is sensitive to those differences. Our simulator enforces this condition through distinct propagation multipliers, and the explanation metrics reveal a benefit that standard onset AUC alone would largely hide.

From a managerial perspective, a typed explanation can distinguish two intervention logics. If a high-risk edge is a primary-supply dependency, mitigation may involve alternate sourcing or inventory; if it is a transport dependency, expediting, rerouting, or carrier substitution is more relevant. RHET-GNN does not prescribe the intervention, but it makes the predicted risk traceable to an entity-relation pair, which is a more actionable object than an undifferentiated attention score.

The favorable properties of the source study [27-32]—directional propagation modeling, intrinsic attention, and explanation evaluation with known cascade paths—remain central. Our contribution is complementary: it shows that once the graph is treated as heterogeneous, explanation quality can improve even when headline predictive metrics barely move. This is a useful caution against selecting explainable risk models solely by AUC.

8. Engineering Deployment Considerations

8.1. Data Contract for a Real Implementation

A practical deployment requires a clear data contract rather than merely a trained graph model. At minimum, the graph layer needs stable entity identifiers, directed dependency endpoints, relation categories, and a dependency-strength proxy. Dynamic features need a timestamped operational signal and exogenous covariates available before the prediction cutoff. Static features should be restricted to attributes that are actually available at scoring time. Missing supplier identities or lane relationships can be represented

explicitly as unknown relations, but silently dropping them would bias both risk estimates and explanations toward the visible portion of the network.

The temporal unit should also match the operational decision cycle. Daily snapshots may be appropriate for many manufacturing settings; transportation control towers may require hourly or event-time updates. RHET-GNN itself does not require a specific unit, but the history window W and the definition of onset must be fixed before model training. A disruption label should represent a decision-relevant event such as production stoppage, severe shortage, or missed service threshold rather than an arbitrary anomaly score, because explanation quality is only meaningful relative to a well-defined outcome.

8.2. From Edge Attribution to Action

At inference time, the system returns a node-level onset probability and a ranked list of incoming typed edges. A decision layer can then group explanations by relation. High primary-supply attribution suggests reviewing alternate suppliers, available inventory, and order prioritization. High transport attribution suggests lane rerouting, carrier substitution, or mode changes. High distribution attribution may motivate downstream allocation or service-level prioritization. Alternate-supply attribution can indicate that resilience capacity itself is becoming unavailable. These mappings are examples of decision interfaces, not prescriptions generated by the model; operational policy should remain governed by business constraints and human review.

An implementation should also expose uncertainty. The current experiments report variability across regenerated seeds, but a production system could estimate uncertainty through ensembles, calibration intervals, or conformal methods. A high-risk score with diffuse attention across many edges is operationally different from the same score concentrated on one dominant supplier. The latter supports targeted intervention; the former suggests a systemic or data-quality issue. Thus, explanation concentration can become an additional triage signal even before causal claims are made.

8.3. Monitoring and Model Governance

Supply networks evolve. New suppliers are approved, lanes change, firms merge, and emergency sourcing can temporarily alter dependency strength. Monitoring should therefore include both predictive drift and graph drift. Predictive drift can be tracked with onset prevalence, calibration, AP, and false-negative rates when labels arrive. Graph drift can be tracked through degree distributions, relation frequencies, edge-weight distributions, and the proportion of nodes with missing upstream visibility. A sudden shift in relation mix can invalidate relation-specific parameters even when node features remain statistically stable.

Explanation monitoring is equally important. Once incident reviews identify likely root causes, top-ranked edges can be compared with reviewed causes to build a real explanation-quality dataset. This creates a path from the simulator-only evaluation used here to field validation. Crucially, the review outcome should be stored separately from model output to avoid circular labels. The system should also log the exact graph snapshot and feature cutoff used for each historical score so that explanations can be reproduced after the network has changed.

8.4. Why the Proposed Model Is Intentionally Lightweight

A more complex heterogeneous temporal transformer could incorporate multiple message-passing layers, continuous-time encodings, and relation-specific multi-head attention. We intentionally avoid that design in the present paper. First, the scientific question is whether relation semantics improve typed attribution under controlled conditions; additional architectural degrees of freedom would make the comparison harder to interpret. Second, supply-chain datasets are often much smaller than web-scale heterogeneous graphs, so parameter efficiency can matter. Third, a lightweight operator makes the complete experiment fast enough to rerun from source and inspect for leakage or hidden preprocessing.

The resulting model should therefore be viewed as a strong engineering baseline rather than a final universal architecture. Its main reusable idea is the separation of three components: temporal encoding of each entity, directed relation-aware propagation, and direct reuse of normalized message weights for explanation. Future work can replace any one component while preserving this decomposition and the same evaluation protocol.

9. Threats To Validity

9.1. Construct Validity

Disruption onset is represented as a binary transition from operational to disrupted. This is appropriate for early-warning evaluation but compresses severity, duration, and partial-capacity loss into one state. Explanation ground truth is normalized incoming contribution at the onset time; other causal definitions could include delayed effects, shared regional shocks, or multi-hop responsibility. We mitigate this limitation by stating the target precisely and evaluating only the target generated by the simulator.

9.2. Internal Validity

Temporal leakage is controlled by chronological splitting and by using only features through time t to predict $t+1$. The F1 threshold is selected on validation rather than test data. Ground-truth edge responsibility is stored by the simulator but is never included in model features or training loss. The same architecture width, optimizer family, feature window, and eligible-node mask are used for both graph models. Randomness is addressed by regenerating the full graph and cascade for three seeds rather than changing weight initialization alone.

9.3. External Validity

The major external-validity threat is the synthetic setting. Real firms may have censored supplier relationships, unobserved inventories, endogenous mitigation actions, and reporting delays. Consequently, the absolute AUC, AP, P@1, and NDCG values in this paper should be interpreted as benchmark outcomes, not expected field performance. A future real-data study should preserve the chronological protocol and should avoid creating pseudo-causal edge labels unless their construction is independently justified.

10. Limitations and Future Work

The most important limitation is external validity. The reported numbers are synthetic and should not be presented

as evidence of performance on a real industrial supply chain. Although the network design, buffers, stochastic shocks, and recovery processes are motivated by the risk-diffusion literature, the simulator remains a stylized abstraction. Real propagation is affected by lead times, bills of materials, capacity substitution, contractual priorities, multi-sourcing shares, inventory policies, and human response.

Second, the graph is static while node states and exogenous signals vary over time. Industrial networks also add and remove suppliers, lanes, products, and contracts. A natural extension is an event-based heterogeneous temporal graph with edge birth/death, possibly using TGN-style memory [33–38] or HTGNN-style cross-time aggregation [18]. Third, attention is evaluated against simulator responsibility but should not be interpreted as a universal causal estimator. Counterfactual edge deletion, intervention-based fidelity tests, and post-hoc explainers such as GNNExplainer or PGExplainer [19], [20] would strengthen future evaluation.

Finally, the real-data opportunity is clear. SupplyGraph [23] provides a useful public starting point for validating temporal heterogeneous representation learning, even though it lacks the causal propagation labels needed here. A pragmatic future study could combine real predictive evaluation on SupplyGraph with synthetic explanation correctness, explicitly separating external predictive validity from causal-attribution validity rather than fabricating unavailable labels.

11. Conclusion

This paper introduced RHET-GNN, a directed relation-heterogeneous temporal graph neural network for supply-chain disruption-onset prediction with intrinsic typed-edge explanations. In reproducible three-seed synthetic experiments, graph modeling raised onset AP from 0.141 for a temporal-only GRU to 0.252 for RHET-GNN. Relation typing provided only a small additional predictive gain over homogeneous attention, but improved top-cause explanation P@1 by 4.4 percentage points and NDCG by 2.5 points. The main empirical lesson is therefore nuanced: topology is decisive for prediction, while heterogeneity is especially important for explaining which operational dependency carried the risk. All metrics, figures, and testbed statistics are generated by the accompanying executable code.

References

- [1] Ren, S., Jin, J., Niu, G., & Liu, Y. (2025). ARCS: Adaptive reinforcement learning framework for automated cybersecurity incident response strategy optimization. *Applied Sciences*, 15(2), 951.
- [2] Li, P., Ren, S., Zhang, Q., Wang, X., & Liu, Y. (2024). Think4SCND: Reinforcement learning with thinking model for dynamic supply chain network design. *IEEE Access*, 12, 195974–195985.
- [3] Hu, X., Zhao, X., Wang, J., & Yang, Y. (2025). Information-theoretic multi-scale geometric pre-training for enhanced molecular property prediction. *PLOS ONE*, 20(10), e0332640.
- [4] Wang, M., Zhang, X., Yang, Y., & Wang, J. (2025). Explainable machine learning in risk management: Balancing accuracy and interpretability. *Journal of Financial Risk Management*, 14(3), 185–198.
- [5] Zhang, H. (2024). Graph-aware multi-task learning for early prediction of timing violations and routing congestion in digital IC physical design. *World Journal of Information Technology*, 2(1), 13–20.

- [6] Wang, J., Tan, Y., Jiang, B., Wu, B., & Liu, W. (2025). Dynamic marketing uplift modeling: A symmetry-preserving framework integrating causal forests with deep reinforcement learning for personalized intervention strategies. *Symmetry*, 17(4), 610.
- [7] Zhao, X., Liu, J., Wang, Y., & Wang, J. (2026). CryptoMamba-SSM: Linear complexity state space models for cryptocurrency volatility prediction. *IEEE Open Journal of the Computer Society*, 7, 226–243.
- [8] Wang, J., Liu, J., Zheng, W., & Ge, Y. (2025). Temporal heterogeneous graph contrastive learning for fraud detection in credit card transactions. *IEEE Access*.
- [9] Sun, T., Yang, J., Li, J., Chen, J., Liu, M., Fan, L., & Wang, X. (2024). Enhancing auto insurance risk evaluation with transformer and SHAP. *IEEE Access*, 12, 116546–116557.
- [10] Zhang, H. (2025). Reinforcement learning approaches for layout optimization in electronic design automation with electromagnetic compatibility constraints. *Frontiers in Robotics and Automation*, 2(2), 77–93.
- [11] Chen, Z., Liu, J., & Chen, J. (2025). Machine learning methods for financial forecasting in enterprise planning: Transitioning from rule-based models to predictive analytics. *Frontiers in Artificial Intelligence Research*, 2(3), 541–564.
- [12] Chen, J., Liu, J., Liang, Y., & Zhou, M. (2026). KE-MLLM: A knowledge-enhanced multi-sensor learning framework for explainable fake review detection. *Applied Sciences*, 16(6), 2909.
- [13] Yue, X., Yang, J., & Liu, W. (2026). FraudDebate-Agent: A multi-agent LLM framework with an evidence-based debate mechanism for financial statement fraud detection. *Mathematics*, 14(15), 2695.
- [14] Zhu, R., & Yue, X. (2024). An explainable machine learning framework for predicting dividend sustainability and financial risk of U.S. REITs under interest-rate shocks. *Journal of Trends in Finance and Economics*, 1(2), 48–57.
- [15] Han, X., Yang, Y., Chen, J., Wang, M., & Zhou, M. (2025). Symmetry-aware credit risk modeling: A deep learning framework exploiting financial data balance and invariance. *Symmetry*, 17(3), 341.
- [16] Chen, J., Cui, Y., Zhang, X., Yang, J., & Zhou, M. (2024). Temporal convolutional network for carbon tax projection: A data-driven approach. *Applied Sciences*, 14(20), 9213.
- [17] Sun, T., Wang, M., & Chen, J. (2025). Leveraging machine learning for tax fraud detection and risk scoring in corporate filings. *Asian Business Research Journal*, 10(11), 1–13.
- [18] Chen, J., Wang, M., & Sun, T. (2025). Intelligent tax systems and the role of natural language processing in regulatory interpretation. *American Journal of Machine Learning*, 6(4), 74–94.
- [19] Yue, X., & Zhu, R. (2024). An explainable hybrid machine learning framework for cash-flow-conditioned multi-horizon liquidity risk early warning: An SME-oriented benchmark study. *World Journal of Management Science*, 2(1), 63–72.
- [20] Zi, B. (2024). Cloud-native distributed systems for real-time payment intelligence. *AI and Data Science Journal*, 1(1), 51–56.
- [21] Zhang, S., & Qiu, L. (2023). An interpretable, uncertainty-aware framework for bridge deterioration prediction and risk-based maintenance prioritization. *Academic Journal of Architecture and Civil Engineering*, 1(2), 21–28.
- [22] Jin, J., Xing, S., Ji, E., & Liu, W. (2025). Xgate: Explainable reinforcement learning for transparent and trustworthy API traffic management in IoT sensor networks. *Sensors*, 25(7), 2183.
- [23] Xing, S., Wang, Y., & Liu, W. (2025). Multi-dimensional anomaly detection and fault localization in microservice architectures: A dual-channel deep learning approach with causal inference for intelligent sensing. *Sensors*, 25(11), 3396.
- [24] Xing, S., & Wang, Y. (2025). Proactive data placement in heterogeneous storage systems via predictive multi-objective reinforcement learning. *IEEE Access*, 13, 117986–117998.
- [25] Ji, E., Wang, Y., Xing, S., & Jin, J. (2025). Hierarchical reinforcement learning for energy-efficient API traffic optimization in large-scale advertising systems. *IEEE Access*.
- [26] Xing, S., & Wang, Y. (2025). Cross-modal attention networks for multi-modal anomaly detection in system software. *IEEE Open Journal of the Computer Society*.
- [27] Liu, Y., Ren, S., Wang, X., & Zhou, M. (2024). Temporal logical attention network for log-based anomaly detection in distributed systems. *Sensors*, 24(24), 7949.
- [28] Yang, J., Li, P., Cui, Y., Han, X., & Zhou, M. (2025). Multi-sensor temporal fusion transformer for stock performance prediction: An adaptive Sharpe ratio approach. *Sensors*, 25(3), 976.
- [29] Zhao, X., Sun, T., Ren, S., Yang, J., & Liu, Y. (2025). RAG-Based AI Agents for Enterprise Software Development: Implementation Patterns and Production Deployment. *Frontiers in Artificial Intelligence Research*, 2(3), 501–520.
- [30] Sun, T., Wang, M., & Chen, J. (2025). Leveraging machine learning for tax fraud detection and risk scoring in corporate filings. *Asian Business Research Journal*, 10(11), 1–13.
- [31] Bian, M., Li, P., Lin, Y., Wang, Y., & Teng, D. (2026). Assessing green public supply chains resilience under carbon neutrality goals: A multi-agent reinforcement learning framework. *IEEE Access*.
- [32] Teng, D. (2025). TEAS: Token-and energy-aware autoscaling for cost-efficient LLM serving. *AI and Data Science Journal*, 6(3), 1–10.
- [33] Rhee, M., Zou, J., Mo, T., Teng, D., & Yang, J. S. (2026). Enhancing web search agents with self-play contrastive fine-tuning. *IEEE Access*.
- [34] Wang, B., Zhao, W., & Wang, Z. (2024). DAS-GNN: A scalable disagreement-aware graph neural framework for anomaly detection in large-scale networks. *AI and Data Science Journal*, 1(1), 67–75.
- [35] Wang, Z., & Shen, Z. (2024). Flex-Talent: An explainable and fair machine learning framework for high-stakes leadership-pipeline evaluation. *World Journal of Management Science*, 2(1), 73–82.
- [36] Guo, Z., Chen, T., & Shang, W. (2022). Class-conditional intermediate-domain adversarial adaptation for cross-device acoustic scene classification. *Journal of Computer Science and Electrical Engineering*, 4(1), 26–35.
- [37] Wang, Y., Bian, M., & Lin, Y. (2022). An explainable temporal graph neural network for disruption prediction and risk propagation in multi-tier supply chains. *Journal of Computer Science and Electrical Engineering*, 4(1), 17–25.
- [38] Zhao, W., & Wang, B. (2022). CAPA: A prediction-driven autoscaling framework for SLO-aware cloud-native machine learning inference. *Journal of Computer Science and Electrical Engineering*, 4(1), 8–16.