Does the Explanation Drift First? Early Warning of Graph Neural Network Degradation

Anonymous authors
Paper under double-blind review

Abstract

Machine-learning models quietly lose accuracy when data drift occurs, and the labels that would reveal it often arrive late or never. A label-free clue is the model's own explanation: when its reasoning changes, something has shifted. Prior work detects this *explanation shift* in tabular data but does not analyse whether it occurs *before* accuracy drops. This paper focuses on graphs. Lead-time between explanation drift and the model's accuracy is measured on fully controlled, synthetic graph streams. The explanation drifts first, by a median of 15 steps, with no false alarms, and a controlled sweep confirms this reflects a real mechanism, not an artefact. The lead is over accuracy, not over every detector. The effect also holds, albeit more weakly, on a realistic IT infrastructure graph (RCAEval). Explanation drift can therefore serve as a simple, label-free early warning that a graph model should be revalidated and updated to new conditions.

1 Introduction

The relevance of AI models' training may decay over time as the world changes. If unattended, accuracy often falls silently because the labels needed to detect the drop are delayed, costly, or simply absent in deployment (Demšar & Bosnić, 2018). A practical monitor must therefore raise the alarm *before* accuracy has already fallen.

Feature attributions are a natural candidate for such a monitor: they can be computed at any time, with no labels (Yang & Kim, 2019; Al Bared & Nassar, 2021). Recent work tracks their drift as a warning sign (Mougan et al., 2022; 2025; Haug et al., 2022; Duckworth et al., 2021).

Two gaps remain. First, these studies ask whether attributions have shifted, not whether that shift *precedes* the loss of accuracy, the very property that would make it an early warning. Second, they are tabular. On graphs, explanation methods such as PGExplainer (Luo et al., 2020) are studied for stability and faithfulness (Agarwal et al., 2023), but never inverted into a timed signal tied to accuracy. This paper addresses the both gaps and studies on graph streams, whether explanation drift arrives before accuracy degradation, and by how much, a quantity called the *lead-time*.

This question is not merely academic; it is highly useful for IT infrastructure diagnosis. A modern system can be perceived as a graph of services, and graph models are increasingly used for automated root-cause analysis (RCA) (Pham et al., 2025). Yet infrastructure evolves constantly, services are deployed, scaled, and re-wired, so a trained model slowly goes stale, becoming less useful for their operators (Szandała, 2025b). Because faults propagate along dependency edges, and resilient systems mask a failing path with redundancy until it saturates, the model's explanation should shift before its diagnoses fail. Explanation drift then becomes an operational trigger: revalidate the RCA model before it silently misdiagnoses.

The contributions are as follows.

- The *precedence* question is stated for GNNs equipped with parametrised edge explainers, and lead-time is defined as the gap between the onset of explanation drift and the onset of accuracy degradation.

- Using synthetic graph streams with a known drift onset, a positive and significant lead-time is demonstrated at a zero-false-alarm operating point, and its connection to a staged-redundancy mechanism is established through a falsifiable dose–response sweep.
- The effect is further affirmed on a realistic IT infrastructure substrate built on the RCAEval service graph (Pham et al., 2025), where precedence persists, albeit to a smaller extent, and the reasons for this difference are explained.
- Comparisons with label-free drift-detection baselines are provided and reported in full, rather than being limited to the most favourable baselines.

2 Related Works

Attributions drift when distributions move. Mougan et al. (Mougan et al., 2022; 2025) formalise this as an *explanation shift*, detecting it from SHAP distributions. Yet they state plainly that their method detects a shift but does not predict performance loss. Also, the study focuses solely on tabular data. The wider monitoring literature (Duckworth et al., 2021; Shanbhag et al., 2021; Haug et al., 2022; LOPATECKI et al., 2024) likewise treats attributions as a label-free alarm. None poses our temporal question: does the alarm sound before accuracy falls, and if so, by how much? And none ventures onto graphs.

A second strand focuses on explaining the drift rather than just anticipating it. Hinder et al. (Hinder et al., 2023) explain drift after the fact, classifying pre- against post-drift segments. This is useful, but it works retrospectively; we should ask whether the explanation moves first.

On graphs, the focus has been on explanation quality. PGExplainer (Luo et al., 2020) is inductive: train it once, then track masks cheaply over time. Agarwal et al. (Agarwal et al., 2023) formalise stability and find PGExplainer the steadiest of the field, while others show that GNN explanations buckle under shift. This hands us both a mechanism and a tool, yet stability is treated throughout as a static virtue to be maximised, never inverted into a timed warning tied to accuracy.

The gap is consistent: attribution monitoring is tabular and detection-only, while graph explanation is studied for stability rather than precedence. The goal of this paper is to test whether explanation drift precedes accuracy degradation with sufficient lead-time.

3 Methodology

3.1 Overview and hypothesis

The goal is to find when the data drifts, does the model's explanation drift before its accuracy falls(as seen in Figure 1)? It is called this temporal gap the lead-time. The hypothesis (H1) is that attribution drift precedes accuracy degradation, and that this lead-time is measurable, robust, and mechanistically explained. H1 is validated on synthetic graph streams with a known, controlled onset of drift, so that both the true cause and the timing are available as ground truth.

3.2 Synthetic data streams

A stream of graph-classification instances is generated over T discrete time steps. The number of nodes n is fixed; no nodes are added or removed. Each step yields a batch of graphs with binary labels. An early reference window $[0, R)$ contains no drift; drift then ramps linearly over $[t_{\text{onset}}, t_{\text{end}}]$.

The primary regime is *redundant structure*. The positive class carries two dense structural motifs, A and B, on disjoint sets of nodes; the negative class carries neither. Node features are uninformative noise, so the only predictive signal is structural and hence visible to an edge-based explainer. Drift degrades motif A *early* and motif B *late*. The intuition is a two-engine system: when A fails, the explanation shifts onto B, yet accuracy holds because B still separates the classes; only when B also fails does accuracy fall. This builds the precedence condition by construction, with a known onset.

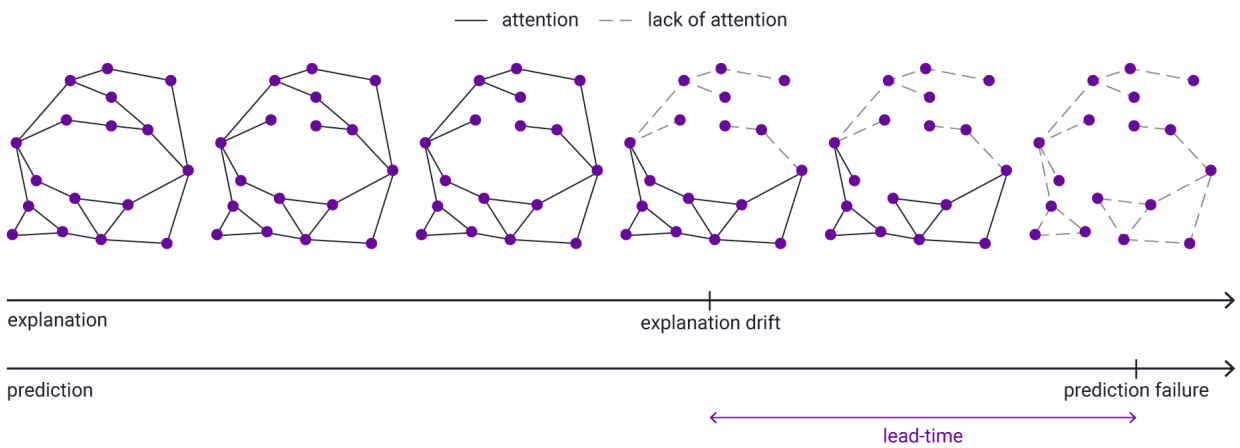

Figure 1: Concept of graph degradation by motiffs and lead-time (source: self)

I use a *no-drift* control (the same generator with drift switched off) to measure false alarms. Likewise, vary the construction (a single, gradually degrading motif and a feature-buffered variant) to check that any effect is not an artefact of a single design.

3.3 Model and explainer

A graph convolutional network is trained for graph classification on the reference window. Next, a parameterised edge explainer (PGExplainer-style; (Luo et al., 2020)) is trained on the same window, with the classifier frozen. The explainer is *inductive*: it is trained once and applied at every subsequent step so that we can cheaply track its edge masks over time, without retraining.

3.4 Signals over time

Each step t emits two signals from a held-out evaluation batch. The first one is Accuracy as well as the error, $e(t) = 1 - \mathrm{acc}(t)$.

Next is the attribution drift, averaging the explainer's edge masks into an aggregate importance vector $\mathbf{a}_t$ over the $n(n-1)/2$ candidate edges. A reference vector $\mathbf{a}_{\mathrm{ref}}$ is the mean over a held-out slice of the reference window. The drift at step t is a distance between $\mathbf{a}_{\mathrm{ref}}$ and $\mathbf{a}_t$.

Most edges carry no signal, so a distance over the full vector is swamped by noise. I use a *focused* drift measure as the primary signal, chosen *a priori* on signal-to-noise grounds, not after seeing results. Let $\mathcal{K}$ be the top-k edges by reference importance, the edges the explainer actually relied on. We can then measure the relative loss of importance mass on those edges(eq. 1),

$$D_{\mathrm{focus}}(t) \;=\; \frac{\sum_{i\in\mathcal{K}} a_{\mathrm{ref},i} - \sum_{i\in\mathcal{K}} a_{t,i}}{\sum_{i\in\mathcal{K}} a_{\mathrm{ref},i}} \tag{1}$$

Full-vector measures (rank correlation, cosine distance) are recorded for transparency.

3.5 Change-point detection and lead-time

Each series is converted to a single onset time using one detector, applied identically to both.

Applied to D_{focus} this gives the attribution onset t_{attr}; applied to $e(t)$ it gives the accuracy onset t_{acc}. The lead-time is (eq. 2)

$$L \;=\; t_{\mathrm{acc}} - t_{\mathrm{attr}}, \tag{2}$$

Table 1: Main result (redundant-structure regime, 10 seeds). Lead-time $L = t_{\text{acc}} - t_{\text{attr}}$; positive means explanation drift came first. The false-alarm rate is measured on 10 no-drift controls at the chosen operating point (λ giving zero control alarms).

Median L	IQR	Frac. $L > 0$	Wilcoxon p	Detection rate	False-alarm rate
15	[9.5, 17]	1.00	$< 10^{-3}$	10/10	0/10

where $L > 0$ means the explanation drifted first. λ stops as the smallest value that yields a zero false-alarm rate on the no-drift controls, and reports the lead-time only at that operating point.

3.6 Experiments

- **Precedence (main).** The primary regime over multiple seeds, with no-drift controls. It asks whether $L > 0$ and whether the detector stays silent without drift.
- **Mechanism (dose-response).** Sweep the degradation *staging* from synchronised (both motifs degrade together) to fully staged (A then B).The goal is for L to grow with staging and fall to zero under synchronisation. This falsifiable boundary ties the lead time to the proposed mechanism rather than to the onset alone.
- **Robustness.** Vary the focus size k, and sweep the detector threshold λ to trace lead-time against false-alarm rate.
- **Generality** I repeat the main test across the alternative regimes to check that the effect is not tied to a particular construction.
- **Baselines (for context).** Finally, run input-feature, embedding and degree-distribution maximum mean discrepancy, and predictive entropy, through the *same* calibrated detector, and compare their lead-times to attribution drift.

3.7 Metrics and goals

Each experiment reports seeds, the median lead-time and its interquartile range, the fraction of seeds with $L > 0$, and a Wilcoxon signed-rank test of $L > 0$. For statistics, also: the detection rate (how often a true drift is caught) and the false-alarm rate (how often the detector fires without drift). The goals are: (i) to show a positive, significant lead-time at a zero-false-alarm operating point; (ii) to show, via the dose–response sweep, that the lead-time follows the staged-redundancy mechanism and vanishes when it should; and (iii) to show the result is stable across k and λ, and consistent across regimes.

4 Results

4.1 Explanation drift precedes accuracy degradation

Across the ten drift seeds, attribution drift was detected in every run, and it was detected before accuracy fell. The median lead-time was 15 steps (IQR $[9.5, 17]$); the lead-time was positive in 10/10 runs, and a Wilcoxon signed-rank test rejected $L \leq 0$ at $p < 10^{-3}$. On the ten no-drift controls, the detector never fired (false-alarm rate 0/10). Thus, the signal is both timely and silent when nothing is wrong. These figures are summarised in Table 1, and a representative run is shown in Figure 2, where the focused attribution-drift curve rises shortly after the drift onset while accuracy holds, and only later collapses.

4.2 The lead-time follows the staged-redundancy mechanism

To test whether the lead-time reflects the proposed mechanism rather than the drift onset alone, the degradation *staging* was swept from synchronised to fully staged (Figure 3). When both motifs degrade together

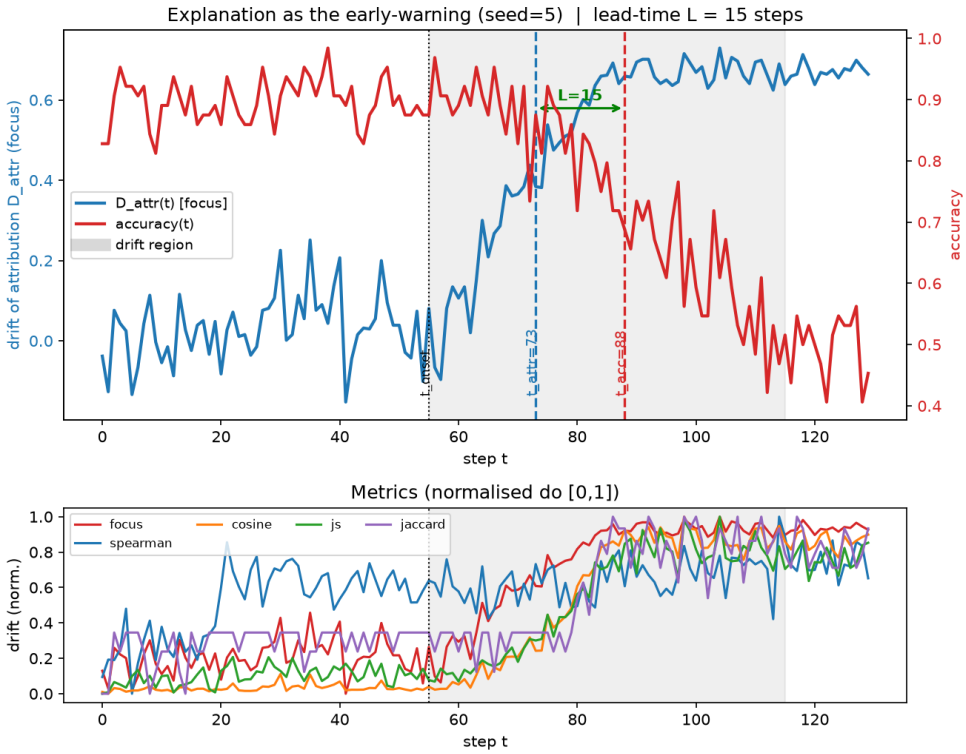

Figure 2: A single run. The focused attribution-drift signal (top, blue) departs from its reference band well before accuracy (top, red) declines; the dashed lines mark the detected onsets t_{attr} and t_{acc}. The lower panel shows the soft detection of each metric.

(synchronised), no precedence was observed: the median lead-time was -5.5 steps and was not significant ($p \approx 0.98$). As staging increased, the lead-time rose monotonically - 0.0, 6.5, 12.0, 15.0 - and became significant from the half-staged point onward ($p < 10^{-3}$). This dose-response behaviour, including the vanishing lead under synchronisation, indicates that the precedence is driven by staged degradation of redundant structure, not by the detector simply firing at onset.

4.3 Robustness

The result was insensitive to the two free choices a sceptic would question. Varying the focus size k over $\{5, 10, 20, 30, 50, 80, 120\}$ (of 300) left the median lead-time essentially flat (between 14.5 and 15.5, all $p \leq 10^{-3}$, false-alarm rate 0); see Figure 4. Sweeping the detector threshold λ traced a clear operating region: for $\lambda \geq 8$ the control false-alarm rate was 0 while the lead-time held between 13 and 15 steps Figure 5). The reported operating point sits inside this plateau.

4.4 Comparison with drift-detection baselines

For context, four label-free baselines were passed through the identical detector and compared on lead-time (Table 2). Attribution drift was a strong leading indicator, but not the earliest: a structural baseline (maximum mean discrepancy on the degree distribution) led by a larger margin (19.5 vs. 15 steps), as expected in a regime where the drift is itself structural. Input-feature discrepancy and predictive uncertainty did not lead at all. Explanation drift can be presented as a *valid* early-warning signal rather than the fastest one. Its distinct value is qualitative - it identifies *which* edges the model has stopped relying on, and it becomes possible to report the baselines in full rather than select among them.

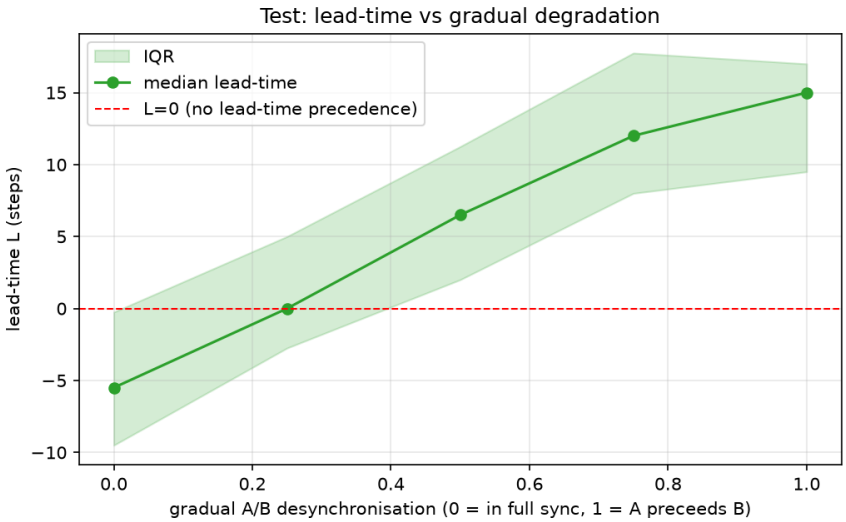

Figure 3: Dose-response. Median lead-time (with IQR band) against degradation staging (0 = synchronised, 1 = fully staged $A \to B$). The lead time increases with staging and vanishes under synchronisation, as the mechanism predicts.

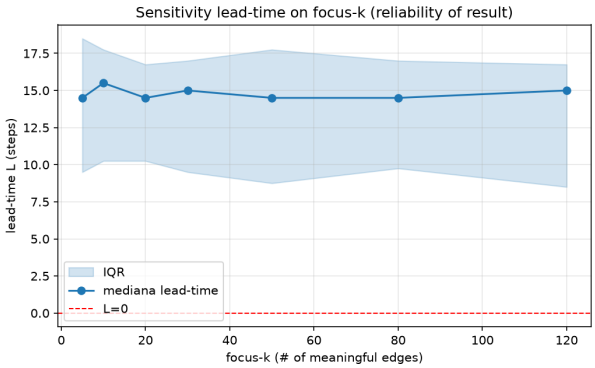

Figure 4: Lead-time is stable across the focus size k.

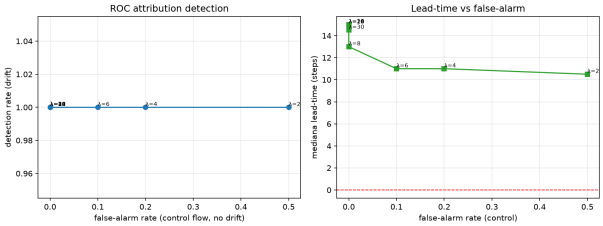

Figure 5: Lead-time and false-alarm rate against the detector threshold λ.

5 Reaffirmation on IT infrastructure (RCAEval)

IT infrastructure always changes, always evolves, thus it is a perfect setting for this explanation drift study (Szandała, 2025a). The dependency graph *is* the system: services are nodes and calls are edges, so faults propagate along the very structure an edge explainer inspects. Resilient deployments also include redundancy, replicas, retries, and failover, which absorb degradation until a component saturates. This is exactly the staged-redundancy mechanism studied on synthetic data: when one path fails, the model's reliance shifts (the explanation drifts), while end-to-end accuracy holds on the backup, and only later collapses. Since infrastructure also develops continuously, a trained root-cause-analysis (RCA) model drifts out of date; asking whether its explanation moves before its diagnostic accuracy falls is therefore operationally useful, as an early signal that the model should be revalidated before it silently misdiagnoses.

This setting was instantiated using a service call graph from RCAEval (Pham et al., 2025), to which controlled, cumulative topology evolution was applied: a semi-synthetic design combining a realistic topology with a fully known onset. The diagnosis task was kept binary (the region of the root cause), so the pipeline of Section Methodology was mostly reused. Precedence was again observed (Figure 6, left): the median

Table 2: Lead-time by signal (same detector, 10 seeds). Reported for context; structural discrepancy leads at least as early as attribution drift in this regime.

Signal	Median L	IQR	Frac. $L > 0$	Wilcoxon p	FA (no-drift)
Attribution drift (ours)	15.0	[9.5, 17]	1.00	$< 10^{-3}$	0/10
Degree-distribution MMD	19.5	[17.5, 21.8]	1.00	$< 10^{-3}$	0/10
Embedding MMD	4.0	[−1.3, 10]	0.70	0.079	2/10
Input-feature MMD	−17.0	[−22.5, −13.5]	0.00	1.00	0/10
Predictive uncertainty	−2.5	[−6.8, 7]	0.40	0.66	0/10

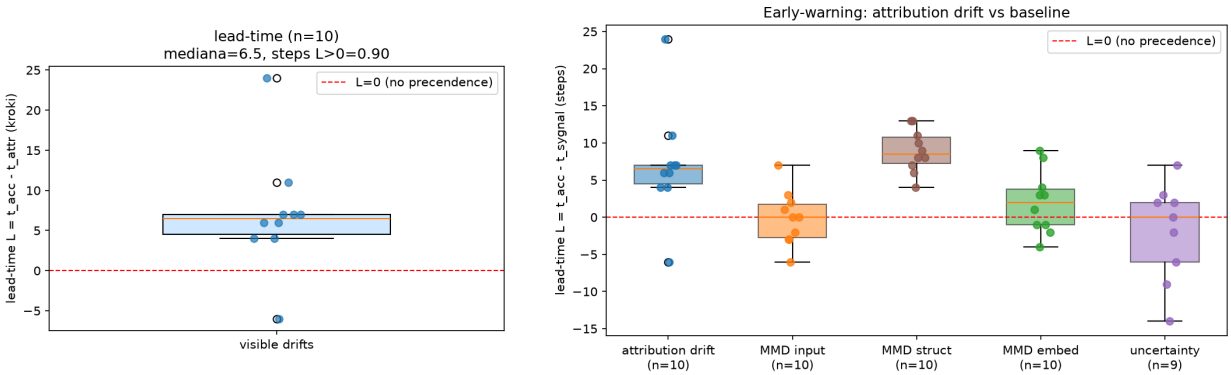

Figure 6: Explanation drift performance on IT infrastructure (RCAEval), 10 drift runs. **Left:** distribution of the attribution-drift lead-time $L = t_{\text{acc}} - t_{\text{attr}}$; the median is 6.5 steps and 90% of runs are positive, so the explanation drifts before diagnostic accuracy falls. **Right:** lead-time by signal, all passed through the same detector. Attribution drift leads clearly to a degree-distribution (structural) baseline leads by a slightly larger margin, while feature, embedding and uncertainty signals do not lead. The lead is smaller and noisier than on synthetic data, as expected for a densely coupled service graph.

lead-time was 6.5 steps, positive in 9/10 runs ($p \approx 0.007$), with no false alarms in the static control. The effect was real but *less pronounced* than on synthetic data (median 15, 10/10). The reason is structural: real service graphs are densely coupled, so the predictive substructure is not isolated, and message passing mixes unrelated changes into the explanation, leaving a noisier signal. Note the wider spread and the occasional negative run in Figure 6 (left). Among the baselines (Figure 6, right), attribution drift was again a clear leading indicator, while a structural discrepancy baseline led by a slightly larger margin and the input, embedding and uncertainty signals sat near zero. Explanation drift thus carries over to IT infrastructure, with a smaller and less clean margin than in the controlled setting.

6 Discussion

A single, falsifiable claim was evaluated: that on graphs, explanation drift precedes accuracy degradation by a measurable lead-time. The evidence supports it. The lead-time is positive, significant, and silent on no-drift controls; it follows the staged-redundancy mechanism through a dose-response sweep, vanishing exactly where the mechanism is absent. This pattern is also observed on a realistic IT-infrastructure substrate. The contribution is therefore narrow but robust; there are several limitations.

The primary result rests on synthetic streams in which the drift onset and the redundant structure are built in by design. This is deliberate, as it provides ground-truth timing but limits external validity. The IT-infrastructure study is semi-synthetic: controlled, cumulative topology evolution applied to a real service graph, not a naturally occurring production incident with an organic onset.

First, the attribution drift leads, but a cheap structural baseline (degree-distribution MMD) leads by a larger margin in both settings. On lead-time alone, the explanation signal does not dominate. Its distinctive value is qualitative; it localises *which* edges or services the model has stopped relying on.

The signal can only see the drift that flows through the substructure the explainer inspects. Drift arriving through a channel the model does not explain through, for instance, a shift in label priors rather than in structure, would move accuracy without first moving the explanation, and would be missed.

On densely coupled graphs, global message passing mixes unrelated changes into the explanation, shrinking and noising the lead, as seen in the smaller RCAEval margin. The method is at its clearest when the predictive substructure is relatively isolated and free of noise.

Only single graph convolutional network and one PGExplainer-style explainer were used. They were chosen due to their SotA position and stability. A noisier explainer or a different architecture could blur or shift the lead. On the other hand, a more robust explainer could be more sensitive to explanation drift.

6.1 Future work

Due to limitations, a few gaps and areas for future work can be identified. The most demanding one is to repeat the study across GNN architectures and explanation methods to determine whether the lead-time is explainer-dependent, and whether more stable explainers yield cleaner leads.

By definition, IT infrastructure is a perfect fit for this finding. Yet it is worth seeking a non-synthetic positive on data with organic structural concept drift and known degradation.

Extend from binary region diagnosis to root-cause *localisation*, and to graphs that grow and shrink as services are added and removed.

Measure whether the drifting subset of edges or services actually points operators to what changed, against logged infrastructure modifications.

It may be worth pairing a structural detector (for *when*) with attribution drift (for *where*, and whether it matters), and studying locality-restricted attributions that resist the coupling effect above. And finally, relate existing explanation-stability bounds to lead-time: under what margin and redundancy conditions must the explanation drift first?

7 Conclusion

Under the evaluated structural degradation conditions, the graph model's explanation drifts before its accuracy falls. On synthetic graph streams, the explanation drift preceded the accuracy degradation by a clear, significant lead time, which vanished exactly when the staged-redundancy mechanism was removed. The effect is carried over more weakly to a realistic IT infrastructure graph.

Explanation drift is not the earliest possible signal; a simple structural baseline is at least as early a signal. But it is label-free and points to *which* part of the model has gone stale. It is a practical early warning that a graph model, such as one used for root-cause analysis, should be revalidated before it silently fails.

References

Chirag Agarwal, Owen Queen, Himabindu Lakkaraju, and Marinka Zitnik. Evaluating explainability for graph neural networks. *Scientific Data*, 10(1):144, 2023.

Doha Al Bared and Mohamed Nassar. Segmentation fault: A cheap defense against adversarial machine learning. In *2021 3rd IEEE Middle East and North Africa COMMunications Conference (MENACOMM)*, pp. 37–42. IEEE, 2021.

Jaka Demšar and Zoran Bosnić. Detecting concept drift in data streams using model explanation. *Expert Systems with Applications*, 92:546–559, 2018.

Christopher Duckworth, Francis P Chmiel, Dan K Burns, Zlatko D Zlatev, Neil M White, Thomas WV Daniels, Michael Kiuber, and Michael J Boniface. Using explainable machine learning to characterise data drift and detect emergent health risks for emergency department admissions during covid-19. *Scientific reports*, 11(1):23017, 2021.

Johannes Haug, Alexander Braun, Stefan Zürn, and Gjergji Kasneci. Change detection for local explainability in evolving data streams. In *Proceedings of the 31st ACM international conference on information & knowledge management*, pp. 706–716, 2022.

Fabian Hinder, Valerie Vaquet, Johannes Brinkrolf, and Barbara Hammer. Model-based explanations of concept drift. *Neurocomputing*, 555:126640, 2023.

Jason LOPATECKI, Aparna Dhinakaran, and Michael Schiff. Data drift impact in a machine learning model, August 6 2024. US Patent 12,056,586.

Dongsheng Luo, Wei Cheng, Dongkuan Xu, Wenchao Yu, Bo Zong, Haifeng Chen, and Xiang Zhang. Parameterized explainer for graph neural network. *Advances in neural information processing systems*, 33: 19620–19631, 2020.

Carlos Mougan, Klaus Broelemann, Gjergji Kasneci, Thanassis Tiropanis, and Steffen Staab. Explanation shift: Detecting distribution shifts on tabular data via the explanation space. *arXiv preprint arXiv:2210.12369*, 2022.

Carlos Mougan, Klaus Broelemann, Gjergji Kasneci, Thanassis Tiropanis, and Steffen Staab. Explanation shift: how did the distribution shift impact the model? *TMLR: Transactions on Machine Learning Research*, 2025.

Luan Pham, Hongyu Zhang, Huong Ha, Flora Salim, and Xiuzhen Zhang. Rcaeval: a benchmark for root cause analysis of microservice systems with telemetry data. In *Companion Proceedings of the ACM on Web Conference 2025*, pp. 777–780, 2025.

Aalok Shanbhag, Avijit Ghosh, and Josh Rubin. Unified shapley framework to explain prediction drift. *arXiv preprint arXiv:2102.07862*, 2021.

Tomasz Szandała. Aiops for reliability: Evaluating large language models for automated root cause analysis in chaos engineering. In *International Conference on Computational Science*, pp. 323–336. Springer, 2025a.

Tomasz Szandała. Chatgpt vs human expertise in the context of it recruitment. *Expert Systems with Applications*, 264:125868, 2025b.

Mengjiao Yang and Been Kim. Benchmarking attribution methods with relative feature importance. *arXiv preprint arXiv:1907.09701*, 2019.