# Toward Explainable GNN-Based Dynamic Stability Assessment via Joint Feature Attribution

Martin Sadric
Karlsruhe Institute of Technology
Karlsruhe, Germany
martin.sadric@kit.edu

Benjamin Schäfer
Karlsruhe Institute of Technology
Karlsruhe, Germany
benjamin.schaefer@kit.edu

## Abstract

Graph Neural Networks (GNNs) offer a fast approach to dynamic stability assessment of power grids, and edge-conditioned architectures are especially well suited because they incorporate transmission-line parameters directly into message passing. For deployment, however, power grid operators need explanations that account for all inputs influencing a prediction: bus states, line parameters, and global operating context. Existing post-hoc GNN explanation methods are not designed for this setting: they focus mainly on node- and graph-level classification tasks, provide limited treatment of edge features, and often ignore graph-level features. We propose a joint attribution framework for GNN predictions that decomposes a prediction into node, edge, and graph-level contributions while preserving completeness, establishing a first formal guarantee for trustworthy joint explanations. In this work, we focus on edge-level regression for dynamic stability assessment.

## CCS Concepts

• **Computing methodologies** → **Artificial intelligence**; **Machine learning**; • **Applied computing** → *Physical sciences and engineering*.

## Keywords

Explainable AI, Graph Neural Networks, Power Grid Stability, Dynamic Stability Assessment, Feature Attribution

**ACM Reference Format:**
Martin Sadric and Benjamin Schäfer. 2026. Toward Explainable GNN-Based Dynamic Stability Assessment via Joint Feature Attribution. In *ACM Sustainability Week 2026 (ACM Sustainability Week Companion '26), June 22–25, 2026, Banff, AB, Canada.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3765611.3815370

## 1 Introduction

Assessing the effect of line outages on power-grid stability can be formulated as an edge-level regression problem: for each transmission line, the model predicts a scalar measure of the resulting system impact. This setting is well suited to GNNs with edge-conditioned message passing, such as NNConv [1], which incorporate line parameters directly into message passing.

For deployment, however, accurate predictions are not enough. Power system operators need explanations identifying which buses, lines, and system-level conditions drove a predicted instability. Such explanations support real-time decisions by helping operators understand and justify model predictions in safety-critical settings [2].

We propose a *joint attribution* framework for edge-level regression that explains a single prediction through node-, edge-, and graph-level contributions, with completeness as an initial formal guarantee. This allows one to distinguish whether a predicted instability is driven primarily by endpoint bus conditions, by the faulted line itself, by the surrounding network, or by global operating context, providing decomposed explanations that are interpretable and actionable for power system operators in safety-critical settings.

## 2 Background and Gaps

Post-hoc explainers for GNNs include perturbation-based, gradient-based, and game-theoretic methods. While effective on common benchmarks, they only partially address our setting.

We begin with gradient-based attribution because it is efficient and formally grounded. Integrated Gradients (IG) [5] assigns attribution scores by integrating gradients along a path from a baseline to the observed input, satisfying key axiomatic properties, most notably *completeness*: the attributed contributions sum to the prediction difference. Recent work partially extends this principle to graph data through Graph-Based Integrated Gradients (GB-IG) [4].

In our setting, five gaps remain:
**(G1)** limited treatment of edge features;
**(G2)** frequent omission of graph-level operating features;
**(G3)** emphasis on node- or graph-level predictions rather than edge-level tasks;
**(G4)** evaluation protocols designed mainly for classification rather than regression; and
**(G5)** lack of realistic physics-grounded benchmarks for faithfulness assessment.
**(G6)** limited consideration of computational and operational requirements for deployment in real-time workflows.

Our goal is therefore to extend gradient-based post-hoc explanations to dynamic stability, addressing limitations identified in prior GNN-based work [3].

## 3 Problem Setting and Preliminary Result

We model the power grid as a graph with node features $\mathbf{x}_v$ at each bus (e.g. active power $P$, reactive power $Q$, and voltage magnitude $V$), edge features $\mathbf{e}_{vw}$ on each transmission line (e.g. resistance $R$ and reactance $X$), and a graph-level feature $\mathbf{u}$ representing global operating context (e.g. a power imbalance $\Delta P$). For an outage of line $(i, j)$, the GNN predicts a scalar stability score $\hat{y}_{ij}$.

Our objective is to explain this prediction relative to a baseline operating point. Here, $A_{\mathbf{x}_v}$, $A_{\mathbf{e}_{vw}}$, and $A_{\mathbf{u}}$ denote attribution scores,


ACM ISBN 979-8-4007-2199-1/26/06
https://doi.org/10.1145/3765611.3815370

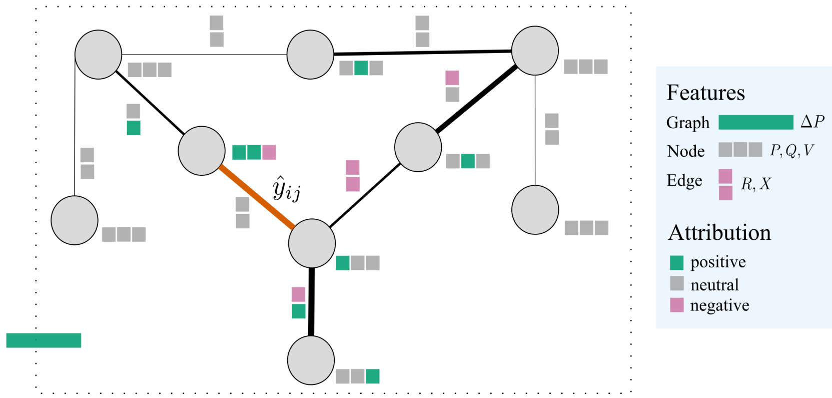


**Figure 1: Joint attribution for an edge-level stability prediction on the faulted line (orange). Contributions are assigned to node features ($P, Q, V$), edge features ($R, X$), and a graph-level operating feature (illustrated here as $\Delta P$) and sum to the prediction difference relative to a baseline operating point.**

i.e. contributions assigned to node, edge, and graph-level inputs toward the prediction difference $\hat{y}_{ij} - y_{ij}^{\text{base}}$:

$$\hat{y}_{ij} - y_{ij}^{\text{base}} = \underbrace{A_{\mathbf{x}_i} + A_{\mathbf{x}_j}}_{\text{endpoint buses}} + \underbrace{A_{\mathbf{e}_{ij}}}_{\text{faulted line}} + \underbrace{A_{\mathcal{N}} + A_{\mathcal{E}_{\mathcal{N}}}}_{\text{surrounding network}} + \underbrace{A_{\mathbf{u}}}_{\text{global context}}, \tag{1}$$

$$y_{ij}^{\text{base}} := F(\mathbf{x}', \mathbf{e}', \mathbf{u}'), \qquad A_{\mathcal{N}} := \sum_{v \neq i,j} A_{\mathbf{x}_v}, \qquad A_{\mathcal{E}_{\mathcal{N}}} := \sum_{(v,w) \neq (i,j)} A_{\mathbf{e}_{vw}}.$$

In practice, we take the baseline as the pre-contingency load flow solution, which is operationally available in any contingency screening workflow. Figure 1 illustrates the decomposition in Eq. (1). We next show that the decomposition in Eq. (1) satisfies completeness.

**Proposition (Grouped completeness).** For a locally Lipschitz, almost-everywhere differentiable scalar-valued GNN $F$, define the straight-line path

$$\gamma(t) = \big(\mathbf{x}' + t(\mathbf{x} - \mathbf{x}'),\ \mathbf{e}' + t(\mathbf{e} - \mathbf{e}'),\ \mathbf{u}' + t(\mathbf{u} - \mathbf{u}')\big), \qquad t \in [0, 1].$$

The grouped attributions are

$$A_{\mathbf{x}_v} := \int_0^1 \nabla_{\mathbf{x}_v} F(\gamma(t)) \cdot (\mathbf{x}_v - \mathbf{x}'_v)\, dt,$$
$$A_{\mathbf{e}_{vw}} := \int_0^1 \nabla_{\mathbf{e}_{vw}} F(\gamma(t)) \cdot (\mathbf{e}_{vw} - \mathbf{e}'_{vw})\, dt,$$
$$A_{\mathbf{u}} := \int_0^1 \nabla_{\mathbf{u}} F(\gamma(t)) \cdot (\mathbf{u} - \mathbf{u}')\, dt.$$

They satisfy

$$\sum_{v \in V} A_{\mathbf{x}_v} + \sum_{(v,w) \in E} A_{\mathbf{e}_{vw}} + A_{\mathbf{u}} = F(\mathbf{x}, \mathbf{e}, \mathbf{u}) - F(\mathbf{x}', \mathbf{e}', \mathbf{u}'). \tag{2}$$

*Proof sketch.* Apply integrated gradients to the grouped input $(\mathbf{x}, \mathbf{e}, \mathbf{u})$ along $\gamma(t)$, then use the chain rule and the fundamental theorem of calculus.

In NNConv, line features enter message passing directly, so the resulting attribution groups map naturally to buses, lines, and global operating context.

## 4 Conclusion and Next Steps

This preliminary result establishes grouped completeness for joint attribution over node-, edge-, and graph-level inputs, demonstrated here for edge-level regression but applicable more generally to node- and graph-level predictions. Remaining challenges are to extend the approach beyond separable message passing and to develop regression-specific faithfulness evaluation, for instance by verifying whether the grouped attributions correspond to physically influential components under dynamic re-simulation. More broadly, the same methodology may extend to other edge-centric domains where edge features carry independent meaning, such as traffic, molecular, and communication networks.

## Acknowledgments

We gratefully acknowledge funding from the Federal Ministry for Economic Affairs and Energy (BMWE, FKZ-03EI1092C) and the Helmholtz Association through Helmholtz AI and the Networking Fund (VH-NG-1727).

## References


[1] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. 2017. Neural Message Passing for Quantum Chemistry. In *Proceedings of the 34th International Conference on Machine Learning*. PMLR, 1263–1272. tex.eventtitle: International Conference on Machine Learning.

[2] R. Machlev, L. Heistrene, M. Perl, K.Y. Levy, J. Belikov, S. Mannor, and Y. Levron. 2022. Explainable Artificial Intelligence (XAI) techniques for energy and power systems: Review, challenges and opportunities. *Energy and AI* 9 (Aug. 2022), 100169. doi:10.1016/j.egyai.2022.100169

[3] Martin Sadric, Sebastian Pütz, Christian Nauck, Veit Hagenmeyer, Frank Hellmann, and Benjamin Schäfer. 2026. Assessing Explanations of Graph Neural Networks for Dynamic Stability Assessment of Power Grids. *ACM Sustainability Week Companion '26, Banff, AB, Canada* (2026). doi:10.1145/3765611.3815508

[4] Lachlan Simpson, Kyle Millar, Adriel Cheng, Cheng-Chew Lim, and Hong Gunn Chew. 2025. Graph-based Integrated Gradients for Explaining Graph Neural Networks. doi:10.48550/ARXIV.2509.07648 Version Number: 1.

[5] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning*. PMLR, Sydney, Australia, 3319–3328. https://proceedings.mlr.press/v70/sundararajan17a.html