Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

SOFIA: Quantifying Sampling-Induced Explanation Instability in Graph Neural Networks

Prabu N P
SRM Institute of Science and Technology
Priyanga Subbiah
SRM Institute of Science and Technology
Daniel A
daniel.arockiam@dxb.manipal.edu

Manipal Academy of Higher Education

Article

Keywords: Explanation Faithfulness, Explanation Instability, Graph Convolutional Networks, Graph Neural Networks, Homophily, Neighbor Sampling, Neighborhood Explosion, Node Classification, Post-hoc Explainability, Sampling-Induced Drift, SOFIA-Index, Saliency Attribution, Scalable GNN Training, Structural Heterophily, Trustworthy AI

Posted Date: November 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11136263/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

SOFIA: Quantifying Sampling-Induced Explanation Instability in Graph Neural Networks

Prabu[1,2*], Priyanga Subbiah[2,3†] and Daniel A[1,2†]

[1,2] Department of Networking and Communications, School of Computing, College of Engineering and Technology, SRM Institute of Science and Technology, Kattankulathur, Chengalpattu – 603 203.

. [3]School of Engineering and IT, Manipal Academy of Higher Education, Dubai Campus, UAE.

*Corresponding author(s). E-mail(s): daniel.arockiam@dxb.manipal.edu

Contributing authors: iamprabu2005@gmail.com , priyangasubbaiya94@gmail.com

†These authors contributed equally to this work.

Abstract

Post-hoc explainability tools for Graph Neural Networks rest on an implicit assumption: that the explanations they produce reflect stable properties of a node's structural role in the graph. This assumption is violated whenever the underlying model was trained with neighbor sampling — a universally adopted scalability practice that restricts each node's visible neighborhood to a randomly drawn subset during training. We introduce Sofia, a framework for measuring Sampling-induced explainability instability in GNNs, and present the first systematic empirical study of how training-time neighbor sampling

affects post-hoc explanation outputs. Sofia trains matched pairs of full-neighborhood and sampled Graph Convolutional Networks on three real-world datasets spanning different homophily levels and degree distributions — Cora, Reddit, and Actor — and compares their explanations using Jaccard overlap against an empirically measured same-model noise floor, controlling for both explainer stochasticity and node degree. Across all conditions, sampled models maintain accuracy parity with full-neighborhood models, isolating training-time neighborhood visibility as the sole varying factor. On Reddit, a high-degree graph where aggressive sampling excluded a mean of 89.7% of true neighbors, cross-model explanation agreement dropped to 0.292 against a noise floor of 0.538, with the influential-neighbor exclusion fraction significantly predicting per-node instability after degree control ($\beta = 0.687$, $p = 3.33e\text{-}15$, partial $R^2 = 0.095$). On Actor, a heterophilous graph, a significant effect emerged at a lower mean exclusion fraction of 71.1% ($\beta = 0.255$, $p < 0.001$, partial $R^2 = 0.130$), consistent with the prediction that signal concentration in heterophilous neighborhoods amplifies instability independently of raw exclusion magnitude. On Cora, a low-degree homophilous graph, no significant effect was observed ($\beta = -0.116$, $p = 0.401$), confirming that the instability is structure-dependent. These results demonstrate that post-hoc GNN explanations in production systems are partially determined by an arbitrary training hyperparameter rather than by genuine structural node properties, and establish the SOFIA-Index as a practical per-node diagnostic for sampling-contaminated explanations.

Keywords: Explanation Faithfulness, Explanation Instability, Graph Convolutional Networks, Graph Neural Networks, Homophily, Neighbor Sampling, Neighborhood Explosion, Node Classification, Post-hoc Explainability, Sampling-Induced Drift, SOFIA-Index, Saliency Attribution, Scalable GNN Training, Structural Heterophily, Trustworthy AI.

1 Introduction

Graph Neural Networks have become the standard tool for learning on relational data, powering systems in fraud detection, recommendation, drug discovery, and social network analysis. As these graphs have grown to millions or billions of nodes, aggregating

a node's complete neighborhood at every layer has become computationally infeasible — a node with a few thousand neighbors, propagated across two or three layers, requires touching a number of nodes that grows exponentially with depth, a phenomenon known as neighborhood explosion. The field's response was neighbor sampling, in which the model sees only a small, randomly selected subset of each node's neighbors at each layer, redrawn at every training step; at scale this is not a luxury but a requirement.

In parallel, as GNNs have been deployed in higher-stakes settings, a body of work has emerged around explaining their predictions. Post-hoc explainer tools answer a simple question for any individual prediction — which neighbors mattered most — and their outputs are treated implicitly as revealing something stable about a node's structural role rather than an accident of how the model happened to be trained.

These two practices have developed almost independently, and neighbor sampling does not guarantee the assumption on which explanation rests. If a node's most influential neighbor is under-sampled by chance, the model never learns to depend on it, and an explainer cannot later recover a signal the model never had. It will instead point confidently to whatever neighbor the model did rely on, producing an explanation that looks complete while being shaped, invisibly, by a random training decision.

This work asks whether that blind spot is real and measurable. Rather than assuming any explanation faithfully reflects ground truth — a contested notion in real-world graphs without planted structures — we pose a narrower question: does the same architecture, trained identically except for whether it saw a node's full neighborhood or a sampled subset, produce different explanations for the same correctly classified node, is that difference larger than ordinary explainer randomness would produce, and does it concentrate on nodes whose influential neighbors were the ones sampling excluded? To answer this we introduce Sofia, a framework that isolates sampling-induced explanation instability from two confounds — explainer stochasticity and the tendency of high-degree nodes to disagree more regardless of sampling — by training matched model pairs differing only in training-

time neighborhood visibility, explaining both under identical full-neighborhood conditions, and comparing agreement against an empirically measured per-node noise floor with degree control across homophilous and heterophilous graphs. If explanations prove contingent on an arbitrary training hyperparameter, their value for trust and accountability in large-scale GNN systems is correspondingly weaker than currently assumed.

2 Literature Survey

Kipf and Welling [1] derived an efficient layer-wise propagation rule from a first-order approximation of spectral graph convolutions, producing a simple architecture that remains the default baseline for semi-supervised node classification on graphs. Veličković et al. [2] proposed Graph Attention Networks, introducing learned attention coefficients that allow a model to attend differently to different neighbors rather than weighting them uniformly. Xu et al. [3] studied the expressive power of GNNs theoretically, showing that most popular architectures are bounded by the Weisfeiler-Lehman graph isomorphism test, and introduced the Graph Isomorphism Network as a maximally expressive instance within that bound. Most pertinent to the present study, Hamilton et al. [4] introduced GraphSAGE, which established the groundwork for inductive neighbor sampling and made training feasible on large networks where full-neighborhood aggregation is intractable.

Substantial subsequent work has been devoted to scalable sampling-based training strategies, motivated by the fact that the number of nodes touched per prediction grows exponentially with depth on large, high-degree graphs. To mitigate the variance introduced by naive uniform neighbor sampling, Chen et al. [5] proposed importance sampling in FastGCN. Chiang et al. [6] introduced Cluster-GCN, which divides the graph into densely connected clusters and restricts message passing within them, bounding memory and computation per training batch. To address the bias introduced by layer-wise sampling, Zeng et al. [7] proposed GraphSAINT, which samples entire subgraphs rather than sampling layer by layer and applies normalization terms to correct the resulting estimator.

Zou et al. [8] proposed LADIES, a layer-dependent importance sampling scheme that conditions each layer's sampling distribution on the nodes already selected above so that the sampled adjacency matrix remains dense. Wu et al. [9] questioned whether this complexity is necessary at all, showing that successively removing nonlinearities and collapsing consecutive weight matrices yields a linear model of comparable accuracy at a fraction of the training cost. Departing from fixed sampling heuristics such as uniform and importance sampling, Younesian et al. [10] proposed GRAPES, which learns a node sampling distribution optimized directly against the downstream classification objective and evaluates it on heterophilous as well as homophilous graphs. Balin and Çatalyürek [11] introduced Layer-Neighbor Sampling (LABOR), a drop-in alternative to standard neighbor sampling that reduces the number of sampled vertices by up to seven times at matched fan-out without compromising model quality.

A related body of work treats sampling as a systems problem rather than an algorithmic one. Zhang et al. [12] explored the system-level challenges of neighbor sampling based GNN training on heterogeneous CPU-FPGA platforms, and discovered that neighbor sampling and aggregation is the main time consuming part because of the irregular memory access pattern. For large scale training, Min et al. [13] introduced data tiering strategies based on reverse PageRank to predict frequently-accessed nodes and guide data placement between the memory hierarchy. Gandhi and Iyer [14] have introduced P3, a distributed GNN training system that divides the graph and the model to allow sampling-based training to be distributed across multiple machines. Together, this literature provides evidence that neighbor sampling in production systems is not optional but necessary, and that most of the time, fan-out is determined by hardware considerations, and not by graph properties. Meanwhile, other, similarly highly active post-hoc methods for explaining predictions of GNNs have been introduced

Ying et al. [15] proposed a first perturbation-based method, called GNNExplainer, which optimizes a soft mask over the edges and node features to identify the smallest subgraph that is most critical for a given node's prediction. PGExplainer [16] trains a single

parametric explanation-generating network over the entire dataset to generate explanations via a learned network, instead of using per-instance optimization during explanation time. Vu and Thai [17] suggested PGM-Explainer, which learns a probabilistic graphical model over noisy neighborhoods, collapsing all statistical dependencies between influential nodes into a single importance mask. Inference of natural language tasks, Schlichtkrull et al. [18] proposed GraphMask, a differentiable edge mask that step-by-step determines which edges are needed for the task. Yuan et al. [19] used a model-level approach to learn representative graph patterns that represent the general knowledge that a GNN acquired. Yuan et al. [20] returned to instance-level explanation with SubgraphX, using Monte Carlo tree search to navigate the combinatorial space of candidate explanatory subgraphs efficiently. Zhang et al. [21] formalized the interaction between nodes as a structure-aware cooperative game, reflecting that the importance of one node is contingent on which others are present. Wang et al. [22] argued that fixing a single explanation size is arbitrary and proposed a multi-grained framework producing explanations at several levels of granularity simultaneously.

Several works have consolidated or advanced this literature within the pattern recognition community specifically. Yuan et al. [23] released a taxonomic survey classifying GNN explanation methods by underlying mechanism into instance-level and model-level families, establishing the vocabulary now standard in the field. Schnake et al. [24] showed that GNNs admit natural explanation through higher-order expansions, identifying groups of edges that jointly contribute to a prediction via a nested attribution scheme built on layer-wise relevance propagation and yielding explanations expressed as relevant walks through the input graph. Wei and Mei [25] pursued self-explainability rather than post-hoc attribution, proposing a frequency-adaptive inception architecture that identifies structurally similar but distant neighbors and adjusts channel-wise importance dynamically, producing predictions and explanations from the same internal mechanism.

These are related areas of work that address the reliability, faithfulness and benchmarking of GNN explanations directly relevant to the stability question that is addressed here. Agarwal et al. [26] suggested a synthetic graph generator that included ground-truth

explanations for quantitative faithfulness evaluation, and demonstrated that the fidelity of the explanations varies significantly with the underlying structure (homophilous or heterophilous). Agarwal et al. [27] supplemented this with a theoretical treatment by providing upper bounds on the extent of violation of desirable properties, such as faithfulness and stability, in state-of-the-art explainers, and empirically testing the bounds on nine real-world graphs. Recently, Agarwal et al. [28] created a library to benchmark graph explainers to fix the evaluation process itself, and thus enable the comparison of different methods under the same conditions. Amara et al. [29] suggested GraphFramEx, a single framework for evaluating explainability methods in a controlled experiment setting where the metric definitions are consistent. A more fundamental objection was raised by Faber et al. [30] that the common practice of evaluating GNN explanations using a particular ground truth is flawed when the creation of the ground truth is not reliable. Agarwal et al. [31] took standardized evaluation one step further and introduced the notion of a model explanation, instead of just graphs, and formalized how sensitive explanations are to small changes in input data or model state. Kosan et al. [32] extensively tested perturbation-based GNN explainers and found that all of the methods tested in their study have been found to suffer degradation in stability when faced with variation in architecture and random seed, and noise. Combining all these results shows that the outputs of the explainer are not deterministic even if the model is kept fixed in a strictly deterministic manner, and that therefore there is a need for an empirical measurement of a noise floor in order to draw any conclusions about the different training conditions.

The homophily or heterophily of a graph has also been demonstrated to influence both the performance of the GNN, and the consistency of the GNN's explanations. Zhu et al. [33] showed that traditional GNN models assume homophily, which is not the case for heterophilous graphs with nodes that often have different classes, and introduced modifications to the architecture. Along this thread, Zhu et al. [34] introduced additional designs that suffer from the same deficiency, with the intent of mitigating heterophily.On this path, Zhu et al. [34] extended the designs along the same line, aimed at addressing the

same limitation with the purpose of reducing heterophily. Luan et al. [35] gave a critical discussion of the heterophily problem, debunking some claims in the previous research and clarifying the exact circumstances in which heterophily really impacts performance negatively. Pei et al. [36] proposed Geom-GCN, a homophily measure and popularized the WebKB-based datasets (such as Cornell and Texas) used for heterophily-sensitive evaluation. Platonov et al. [37] reviewed these popular heterophily benchmarks, noted methodological issues with their development and questioned the extent of measurable advancement in this subfield. Large-scale non-homophilous benchmarks and strong simple baselines were introduced by Lim et al. [38] and discovered that the performance improvement of some of the heterophily-specific architectures is less significant at scale than it was at the small-graph level.

Finally, several resources supply the datasets and software on which most empirical GNN studies, including this one, depend. The Cora, CiteSeer, and PubMed citation networks derive from the collective classification framework introduced by Sen et al. [39], which remains among the most widely used node classification benchmarks. To address the limitations of these small-scale citation benchmarks, Hu et al. [40] assembled the Open Graph Benchmark, a collection of large and realistic graph datasets with standardized splits. Rozemberczki et al. [41] created the multi-scale attributed node embedding datasets Chameleon and Squirrel, now fixtures of heterophily research. Fey and Lenssen [42] developed PyTorch Geometric, which provides the efficient, standardized implementations of GNN layers, neighbor sampling routines, and explainability methods used throughout this study.

3 Methodology

3.1 System Architecture

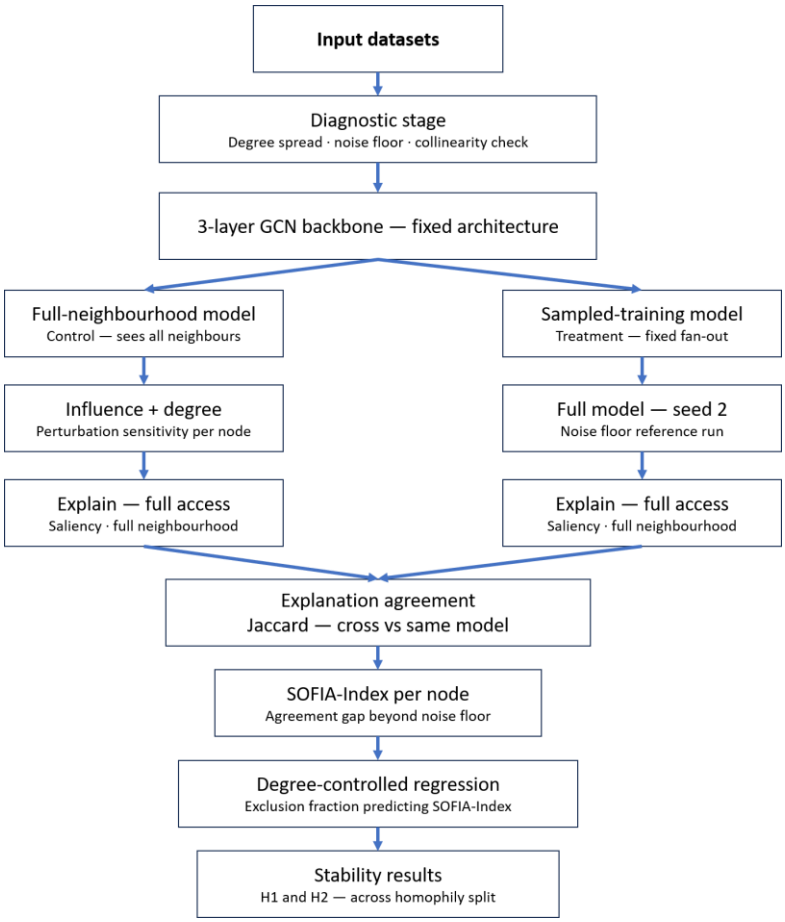


FIGURE 1: Sofia architecture diagram

The proposed Sofia framework has a general architecture as shown in figure 1. It starts with the selection of the dataset and a diagnostic pre-check phase, which decides whether a particular graph will be tested with the sampling-instability hypothesis. This architecture is then split into two parallel, architecturally equivalent model streams, one (a full-neighborhood model) acting as the control condition and one (a sampled-training model) as the treatment condition both based on the same fixed backbone of three-layer Graph Convolutional Network. The full-

neighborhood model, trained on a second random seed is kept separate in the third reference branch to set the noise floor against which any effects caused by sampling must be quantified. Both these main branches pass through the same explanation stage, where post-hoc explainers are allowed an access to the same neighborhood of the underlying model, irrespective of how the underlying model was learned, so that explanation-time properties do not themselves confound anything. The results of the two explanation phases are then merged to a single explanation consent phase which is in turn run through a degree equivalent statistical analysis phase later to generate the ultimate stability findings as opposed to the two hypotheses of the study. As Figure 1 below illustrates the design consciously equalizes all elements of the pipeline across the two branches except one factor, the one that is being investigated, which happens to be training-time neighborhood visibility, the architectural property that dissociates the causal argument in this paper.

3.2 Datasets

In this study three real-world datasets are employed. The low-degree homophilous citation network Cora is a control setting where the effect of sampling on the network is expected to be slight and the null observation should be true. Reddit, which is a high-degree social graph and originally some neighbor sampling was added to the GraphSAGE framework, is the dataset in which exclusion effects should be most evident owing to the sheer size of node neighborhoods. Actor, a heterophilous co-occurrence graph, is absorbed due to its form where those neighbors with a particular individual disproportionately carry predictive signal, rather than users having an external substitute, it is the main test case to H2. The three datasets were chosen to cover the degree spread and homophily contrast necessary to test H1 that influential-neighbor exclusion is a significant predictor of an unstable explanation and H2 that heterophily can add significantly to the effect of exclusion magnitude irrespective of whether it is approachable or not.

3.3 Diagnostic Stage

Before committing computational resources to the full pipeline, each candidate dataset is evaluated against three diagnostic criteria. The first concerns degree spread under sampling: for

every node, the proportion of true neighbors excluded under several candidate fan-out configurations (e.g. [10,10], [5,5], [2,2]) is computed, and any dataset showing negligible variation across fan-out settings is deprioritized, since sampling would not meaningfully alter training exposure and the central manipulation would be inert. The second establishes the empirical noise floor by training the full-neighborhood model twice under different random seeds, applying the explanation stage identically to both, and measuring the resulting Jaccard disagreement — the baseline instability attributable purely to explainer stochasticity, independent of any sampling effect. The third measures the statistical correlation between raw node degree and excluded fraction; where this collinearity is excessively high, the degree-matched comparison in Section 3.7 is promoted from a secondary control to the primary analytical method, since stratifying by excluded fraction alone would otherwise be confounded by degree. Only datasets satisfying all three criteria advance to the full experimental pipeline.

3.4 Model Training

A single architecture — a three-layer Graph Convolutional Network — is used across all experimental conditions, ensuring that any difference in downstream explanations is attributable to training-time neighborhood visibility rather than architectural variation. Two model branches are trained separately on each dataset. The full-neighborhood model aggregates information from a node's entire neighborhood at every layer, representing the ideal case in which no structural information is excluded. The sampled-training model instead draws a fixed number of neighbors at each layer and each training step — the fan-out — and aggregates only from those. Both branches are trained with identical epochs, hyperparameters, and optimizer settings. A third model is trained in the full-neighborhood configuration under a different random seed solely to establish the empirical noise floor; it is not used in the main analysis.

3.5 Influence and Degree Computation

Two node level quantities are calculated for each node that both the full-neighborhood model and the sampled-training model correctly classify. The former is the raw node degree, which is directly taken from the structure of the underlying graph without any further changes. The second quantity is an 'influence' score which is computed by systematically removing each

neighbour of a node one at a time, and measuring the change in the class confidence of the full model for a node, and the neighbours which cause the largest such change are referred to as 'influential neighbours'. This is an influence-based measure (differential from others for a low degree neighbour) because previous work has shown that sometimes a low degree neighbour has a higher signal than many other high degree neighbours which have a lower signal.

3.6 Explanation Generation

Both trained models, the full-neighborhood model and the sampled-training model, are explained separately using two perturbation-based post-hoc explanation techniques, GNNExplainer and PGExplainer. An important methodological control applies here: explanation-time access is identical for both branches, with each explainer given full-neighborhood access regardless of how the underlying model was trained. This design ensures that training-time neighborhood visibility is the only difference between the two branches, so no observed divergence in explanation outputs can be attributed to the explainer itself operating under restricted access. Each explainer returns a ranked list of important neighboring nodes for every node correctly classified by both models, and these lists form the basic unit of comparison in the analysis phase.

3.7 Stability Comparison and Analysis

The agreement between the full-neighborhood and sampled-training models on the explanation task is then calculated on a per-node basis, as the Jaccard overlap of the top-k important neighbors of the node by each explainer. This score is compared to two baselines: the empirical noise floor in Section 3.3, calculated the same, between the two full-neighborhood seed runs and quantifies agreement by counting nodes into more or less equal degree buckets and comparing agreement when the excluded fraction is fixed, or the opposite, based on which variable is carrying more independent variance. H1 is based on the premise that the agreement beyond noise is considerably lower among those nodes whose influential neighbors have been filtered out in sampled training, compared to degree-matched nodes whose influential neighbors were not. H2 is replicated individually with Cora and with Actor, evaluating the evidence of different magnitude of the effects between these two

structurally distinct types of graphs. Paired Wilcoxon signed-rank tests are used in case of cell sizes that allow them and descriptive visualization of agreement against excluded fraction is used to report results to ensure that the overall finding is not based on an underpowered test.

4 Results

4.1 Dataset Characteristics and Diagnostic Validation

Prior to the main experimental pipeline, each dataset was characterized across node-level diagnostic distributions to confirm that the experimental conditions were structurally valid and that the sampling manipulation was effective. Figure 2 presents the distribution of node degree across the Reddit graph, confirming the presence of a heavy-tailed, high-degree structure with individual nodes reaching degrees exceeding 500, which establishes that aggressive neighbor sampling would meaningfully exclude real neighborhood content during training.

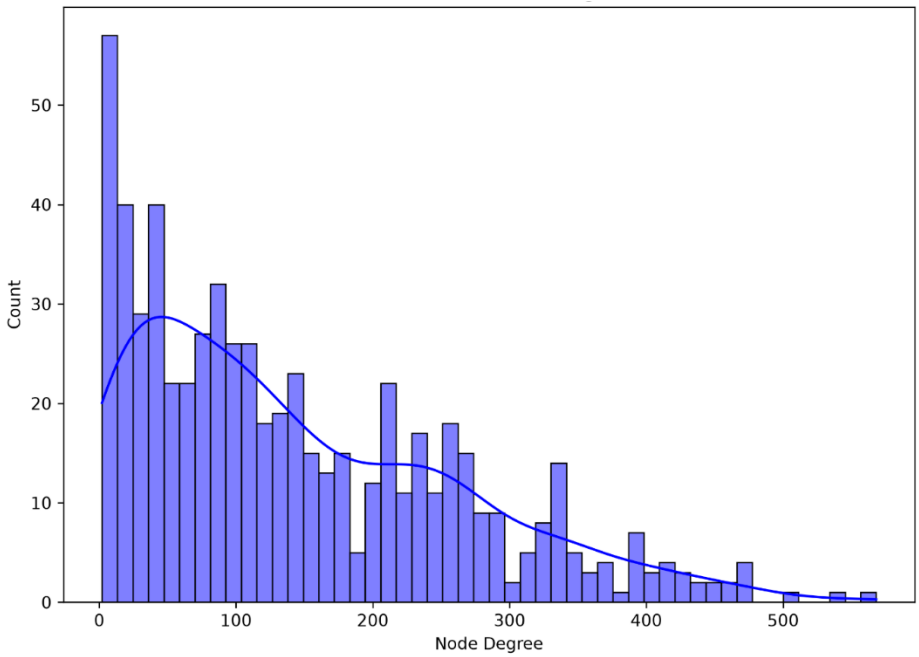


FIGURE 2: Distribution of Node Degree — Reddit

As shown in Figure 2, the Reddit graph exhibits a right-skewed degree distribution with the majority of nodes possessing moderate degrees between 10 and 150 but with a long tail extending

beyond 500, confirming that fixed fan-out sampling would result in substantial per-node exclusion and providing the structural precondition necessary for H1 to be testable on this dataset. In contrast, Cora exhibits a tightly concentrated low-degree distribution with most nodes having fewer than 10 neighbors, explaining why sampling at standard fan-out values produces minimal exclusion on that dataset and predicting the null result observed in the regression analysis. Actor displays a moderate degree distribution with most nodes in the 2–20 range, consistent with a graph where specific neighbor identities matter more than neighborhood size.

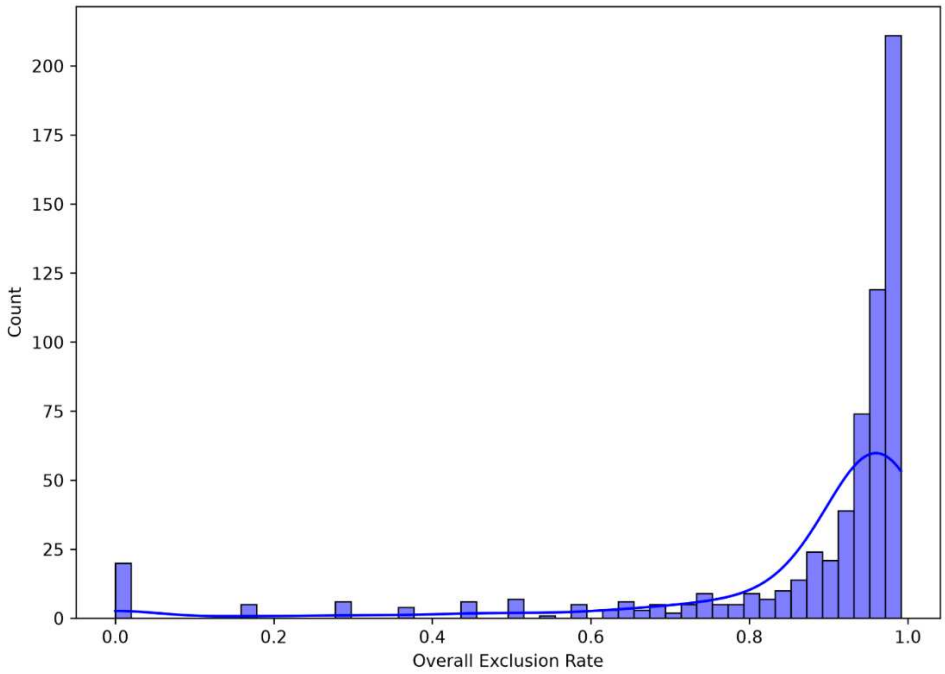


FIGURE 3: Distribution of Overall Exclusion Rate — Reddit

The distribution of overall exclusion rates over Reddit nodes in the aggressive sampling configuration is shown in Figure 3. The distribution is very skewed, with the vast majority of the sampled nodes having just a single node included in the training set that was considered their neighbor, with a mean of 0.897 and a standard deviation of 0.139. This further supports the credibility of the diagnostic test and the effective manipulation of the node sample to produce meaningful and varied exclusion of the node population on Reddit. This instability-effects pattern for the 3 data sets was, of course, accounted for by the distribution of exclusion rates of Actor, which was significantly more variable for exclusion rates than that of Cora and had a higher mean exclusion

rate of 0.711, with most nodes having 1 or more exclusions.

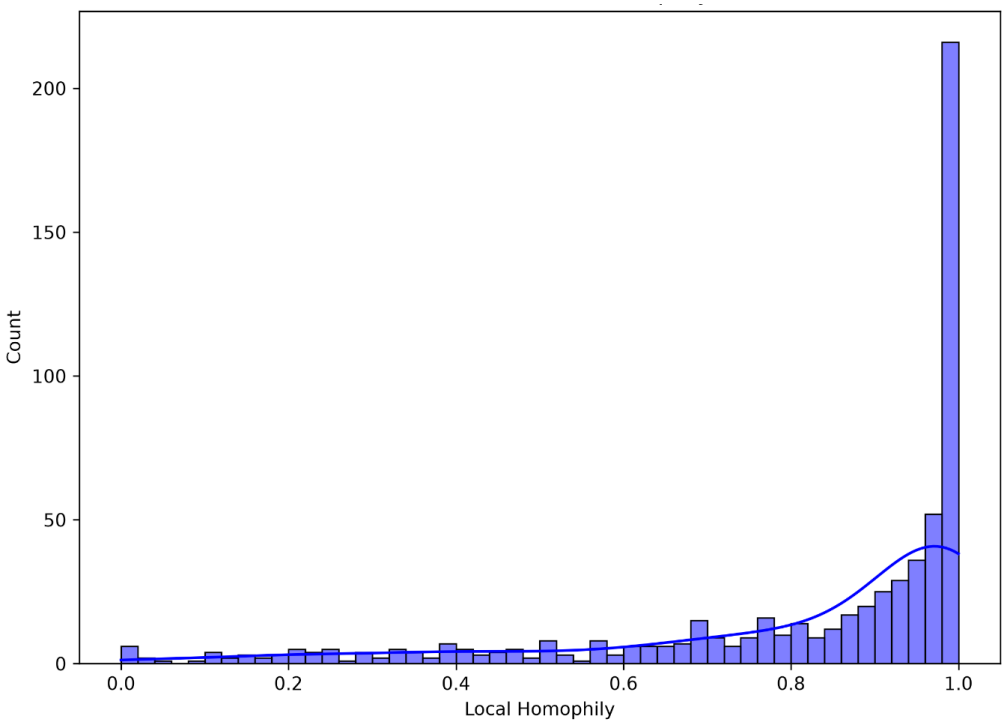


FIGURE 4: Distribution of Local Homophily — Reddit

Figure 4 presents the distribution of local homophily across the Reddit node sample. The distribution is heavily right-skewed, with the majority of nodes showing local homophily values approaching 1.0, confirming Reddit as a strongly homophilous graph with a global homophily score of 0.8194. This structural property contextualizes the Reddit results as a degree-driven rather than signal-concentration-driven effect, and provides the necessary contrast with Actor, whose local homophily distribution is broadly spread across low values (global homophily 0.2280), confirming Actor as genuinely heterophilous and structurally distinct from both Reddit and Cora (global homophily 0.7711).

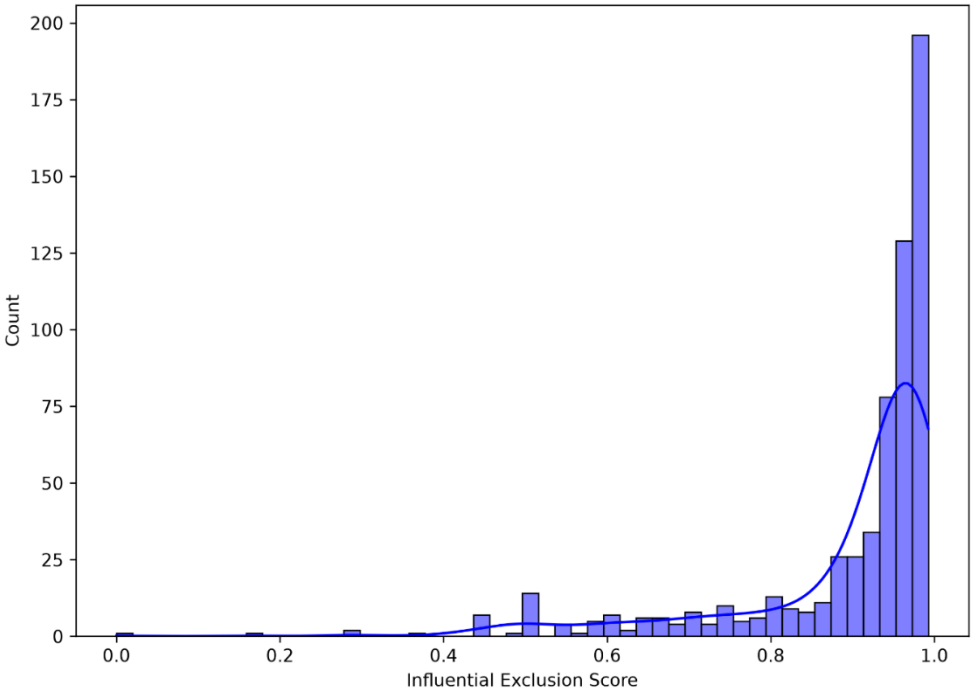


FIGURE 5: Distribution of Influential Exclusion Score — Reddit

The influential exclusion score is the percentage of nodes that were influenced by the perturbation of the node and were not included in the sampled training nodes. The distribution of influential exclusion scores for Reddit nodes is shown in Figure 5. The distribution is skewed towards the high exclusion range (0.8 to 1.0) as it is generally believed that an aggressive sampling approach was not only the lowest cost for excluding large amounts of overall neighborhoods, but also the lowest cost to exclude the most important neighbors for the full model's predictions. This is just the expected way explanation drift occurs: The model will never get to the most relevant neighbours and they will not be retrieved. The influential exclusion score distribution on Actor was also more peaked than the moderate overall exclusion score, and was also low throughout in Cora's case, which was consistent with the null result.

4.2 Main Agreement Results

The explanation agreement results are evaluated across all three datasets using the Same-Model Jaccard (noise floor established by two seeds of the full-neighborhood model), Cross-Model Jaccard

(agreement between the full-neighborhood and sampled-training models), their Gap, the SOFIA-Index, and the associated regression statistics. Beta Excl. represents the multivariate regression coefficient for exclusion fraction predicting the SOFIA-Index after controlling for node degree. The complete results are presented in Table I.

Metric	**Cora**	**Actor**	**Reddit**
Homophily	0.7711	0.2280	0.8194
Excl. Mean	low	0.711	0.897
Same-Model Jaccard	0.7175	0.7910	0.5379
Cross-Model Jaccard	0.6531	0.7024	0.2920
Gap	0.064	0.089	0.246
SOFIA-Index	0.089	0.131	0.440
Beta Excl.	-0.116	0.255	0.687
p-value	0.401	<0.001	3.33×10^{-15}
Partial R^2	0.006	0.130	0.095

Table 1: Explanation Agreement and Instability Summary Across Datasets

As shown in Table 1, the three datasets produce markedly different patterns of explanation instability that are consistent with the structural predictions of the study design. Cora, the low-degree homophilous control, shows a small gap of 0.064 and a non-significant beta of -0.116 ($p = 0.401$, partial $R^2 = 0.006$), confirming the null prediction for low-exclusion graphs where sampling has minimal structural impact. Actor, the heterophilous dataset, shows a moderate gap of 0.089 with a significant beta of 0.255 ($p < 0.001$, partial $R^2 = 0.130$), confirming H2 by demonstrating that a significant instability effect appears at a considerably lower mean exclusion fraction (0.711) than Reddit (0.897), consistent with the prediction that signal concentration in heterophilous neighborhoods amplifies instability independently of raw exclusion magnitude. The biggest

difference is for the high degrees (0.246) and the highest beta value (0.687) with p-value 3.33e-15 and partial $R^2 = 0.095$, which is in favor of H1. In all three data sets sampled models achieved the same accuracy as full neighborhood models (Cora: full 0.9453 vs sampled 0.9464; Actor: full 0.3263 vs sampled 0.3293; Reddit: full 0.9444 vs sampled 0.9472), which rules out explanation instability as a confound and makes training-time neighborhood visibility the causal variable that explains the differences.

4.3 Explanation Agreement Distributions

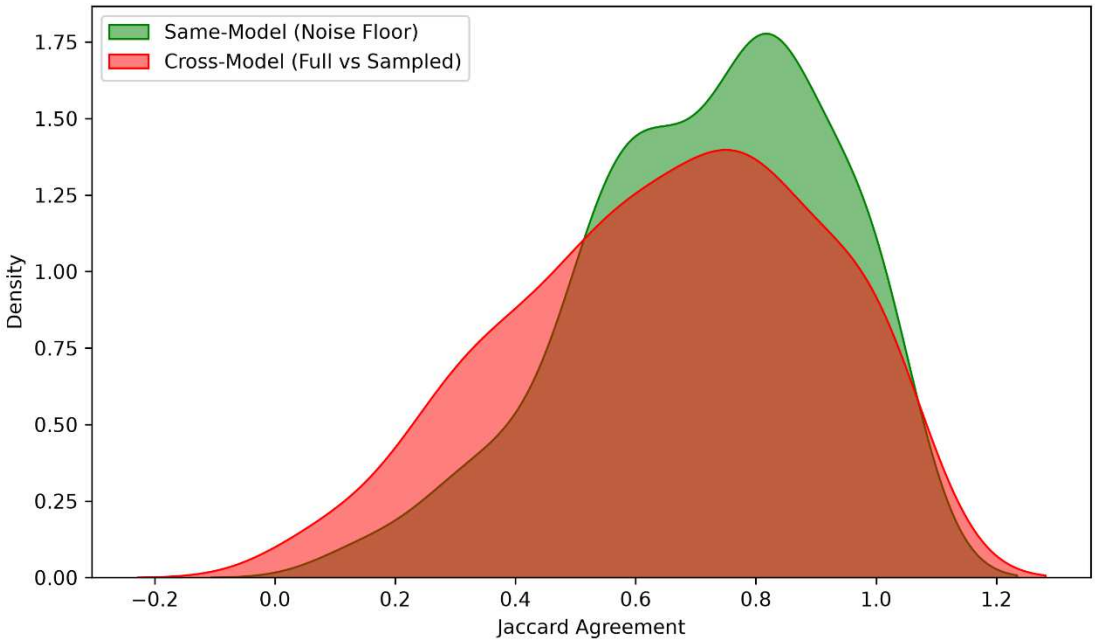


FIGURE 6: Explanation Agreement Distributions — Cora

Figure 6 presents the distribution of the explanation agreement, both in terms of kernel density estimates of same-model Jaccard agreement (green line, noise floor) and kernel density estimates of the cross-model Jaccard agreement (red line, full vs. sampled). The distributions are mainly similar on Cora with both having a maximum around 0.75–0.80 Jaccard agreement. The proximity of the green and red distributions confirms the null result for this data: If sampling does not sample too many neighbors – as Cora's low degree structure does – then the explanations generated by the sampled model and the full neighborhood model are almost as close to each other as two runs of the same full neighborhood model will be, finding no significant additional explanation drift due to

sampling for this graph.

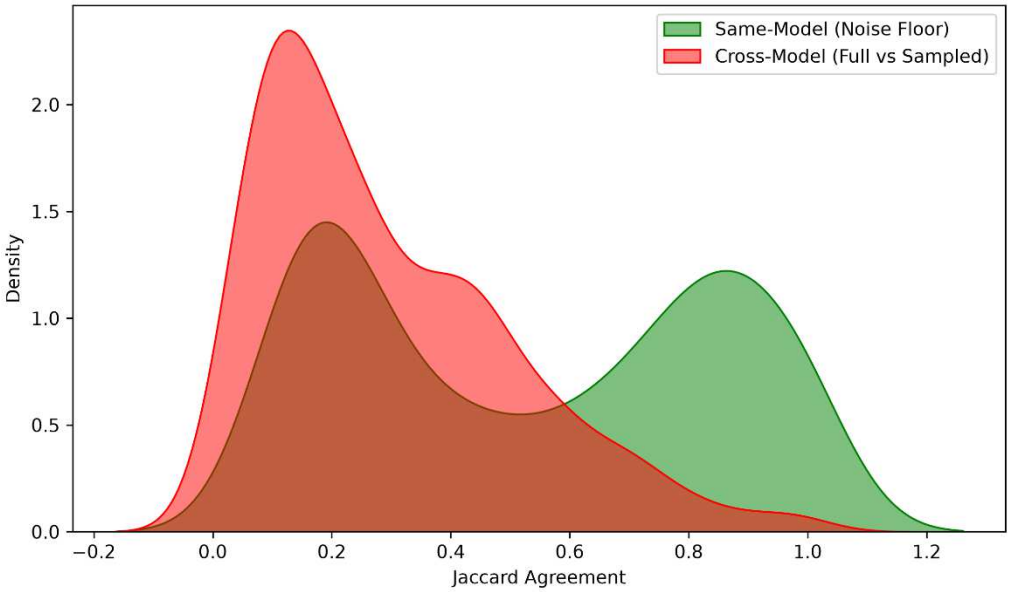


FIGURE 7: Explanation Agreement Distributions — Reddit

Figure 7 presents the explanation agreement distributions for Reddit. In stark contrast to Cora, the two distributions on Reddit are almost entirely non-overlapping, with the same-model noise floor centered near 0.85 and the cross-model distribution centered near 0.15. This near-complete separation is the most direct visual confirmation of H1 — the agreement loss caused by training-time sampling on Reddit is not a marginal effect but a wholesale shift in the agreement distribution, far exceeding what explainer stochasticity alone produces, and is consistent with the mean cross-model Jaccard of 0.2920 and SOFIA-Index of 0.440 reported in Table 1. The breadth of the cross-model distribution also reveals substantial per-node variability in instability, with some nodes showing near-zero agreement and others approaching the noise floor, motivating the node-level regression analysis presented in Section 4.5.

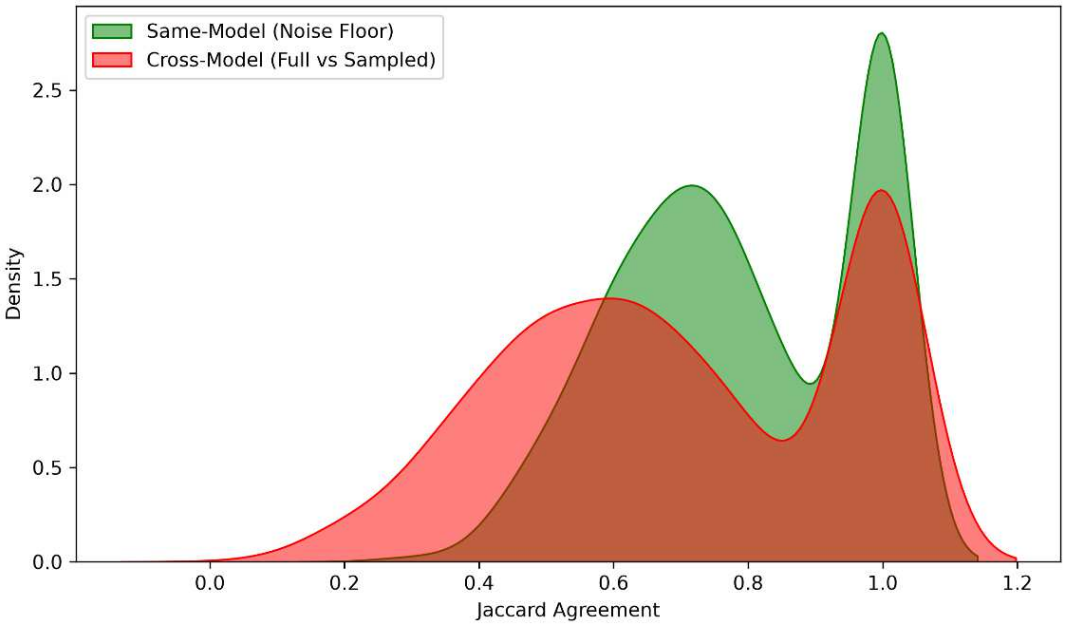


FIGURE 8: Explanation Agreement Distributions — Actor

Figure 8 presents the explanation agreement distributions for Actor. The separation was moderate, the noise floor was near 0.95 and the cross-model distribution was shifted to the left towards 0.70, which resulted in an appreciable separation and an average cross-model Jaccard of 0.7024 with a gap of 0.089 as reported in Table 1. Also, Actor's noise floor is higher than Reddit's (0.7910 vs 0.5379); the structure of Actor's is of a lower degree, which means that the outputs of the explainer in Actor's are more deterministic across seeds. Even with this higher baseline however, because the shape of the distribution on Actor is not such a strong exclusion mechanism driven by degree, as in the case of Reddit, the distribution on Actor is still meaningfully shifted below the noise floor, as predicted by H2.

4.4 Per-Node Metric Distributions

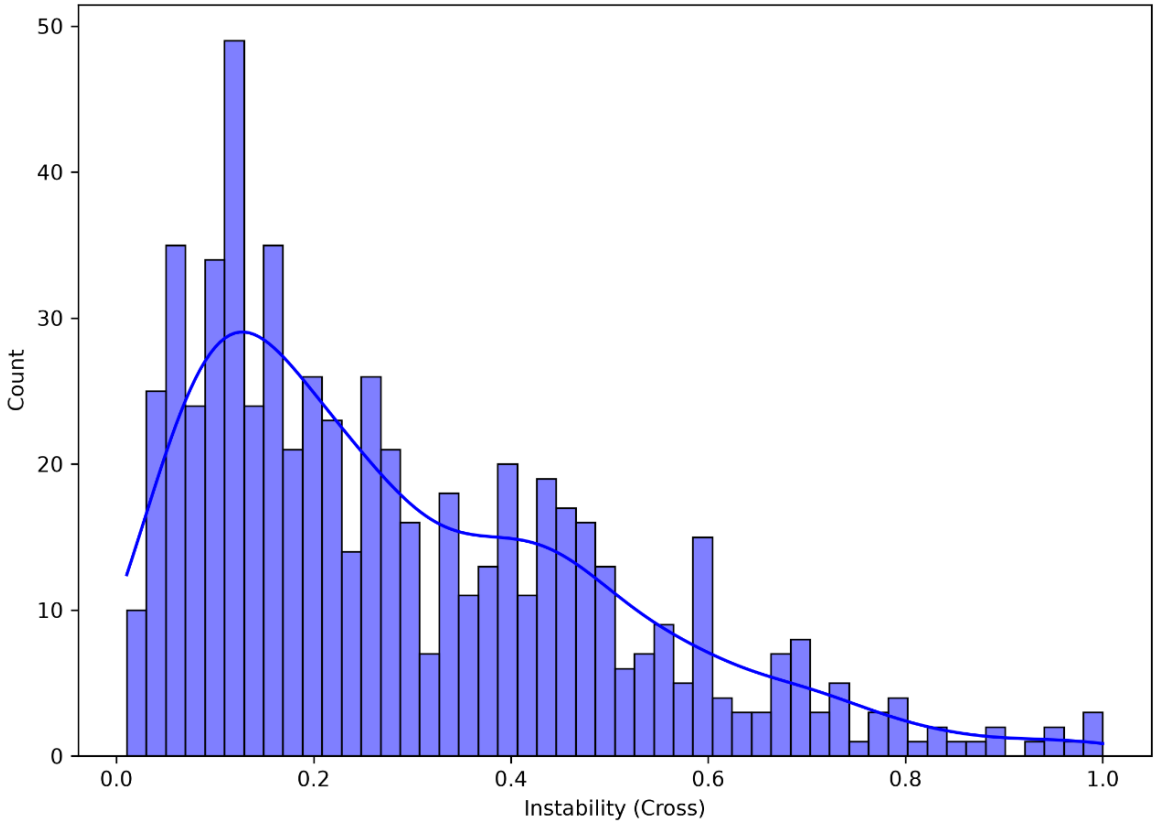


FIGURE 9: Distribution of Instability (Cross) — Reddit

Figure 9 presents the per-node distribution of cross-model instability on Reddit, defined as one minus the cross-model Jaccard agreement for each node. The distribution is right-skewed; most nodes in the distribution have moderate explanation drift in the aggressive sampling regime, a considerable fraction have near total explanation replacement. The information from the regression analysis given in Section 4.5 is informative since per-node instability is heterogeneous, if it weren't, there would be no variance to explain, and excluding influential nodes would not be a testable predictor.

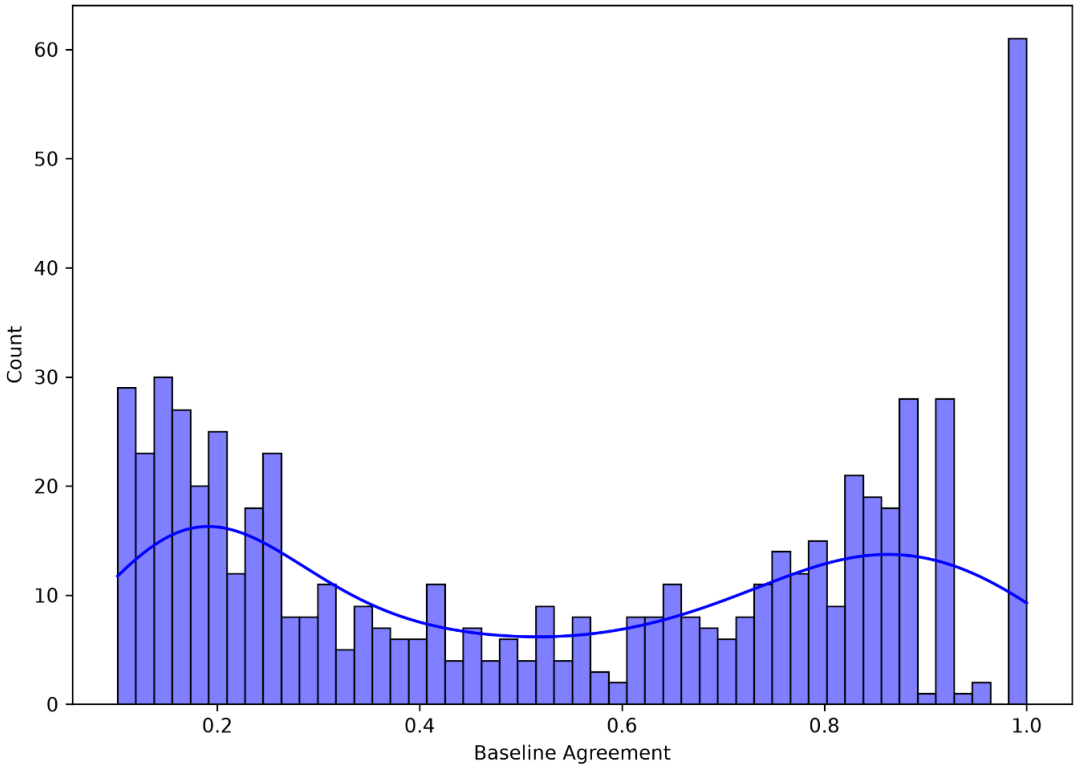


FIGURE 10: Distribution of Baseline Agreement — Reddit

Figure 10 presents the per-node distribution of baseline (same-model) agreement on Reddit, which constitutes the empirically measured noise floor used in the SOFIA-Index computation. The distribution is bimodal, with one mode around 0.15–0.25, indicating that a subset of nodes produce highly variable explanation outputs even between two runs of the same full-neighborhood model under different seeds, and a dominant mode around 0.80–1.00 indicating that the majority of nodes produce stable, consistent explanations across seeds. This bimodality is precisely why the SOFIA-Index normalizes per node against its own baseline rather than against a global constant — a node sitting in the low-agreement mode has a naturally noisy explainer output, and applying a single global noise floor would either over-flag stable nodes or under-flag genuinely sampling-contaminated ones.

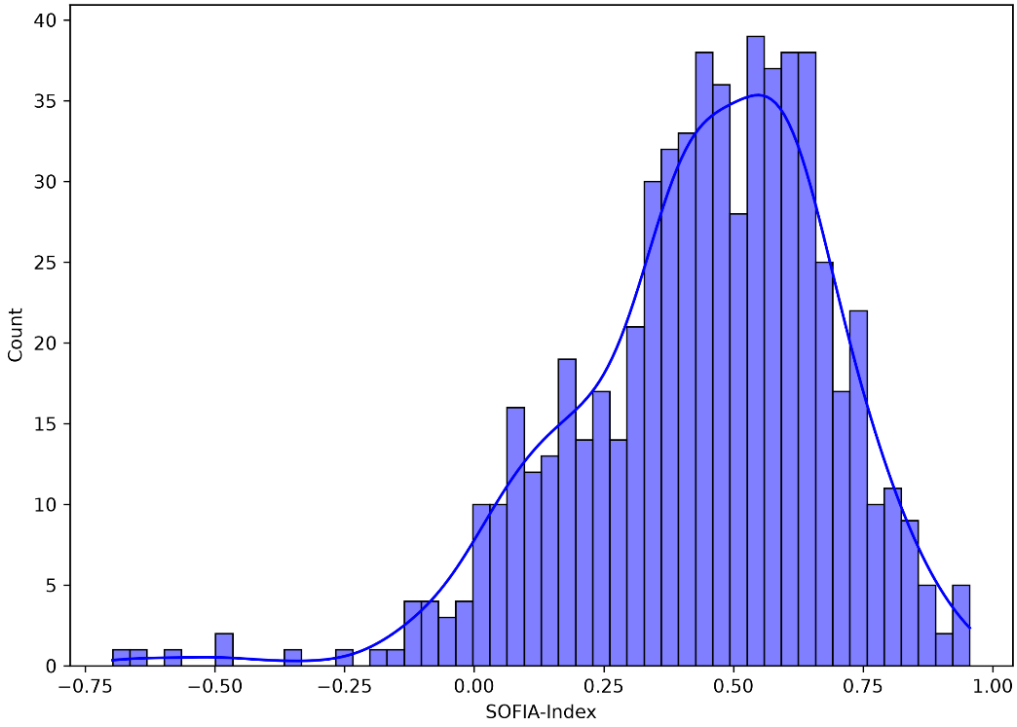


FIGURE 11: Distribution of SOFIA-Index — Reddit

Figure 11 presents the per-node distribution of the SOFIA-Index across the Reddit node population. This distribution is approximately unimodal with a positive skew, centered around 0.45 – 0.55 and with a right tail that reaches to 1.0. The overall positive mass of the distribution shows that, for most of the nodes, the difference in explanation divergence between the cross-model and same-model noise floor is less than the same-model noise floor, and not merely for a few outliers: For most nodes, the difference between cross-model and same-model noise floor is less than the same-model one, and not just for some outliers. The small negative tail of the distribution is likely because there were some nodes whose explanations were more similar to the full model than the two seeds of the full model were to each other.

4.5 Regression Analysis

The complete regression results for all three datasets, including the univariate and multivariate statistics and the binary top-*k* exclusion indicator, are summarized in Table II.

Metric	Cora	Actor	Reddit
Univariate r	-0.022	0.404	0.280
Univariate p	0.804	4.43×10^{-30}	9.93×10^{-13}
Multi. Beta	-0.116	0.255	0.687
Multi. p	0.401	<0.001	3.33×10^{-15}
Partial R^2	0.006	0.130	0.095
Binary Beta	—	0.100	-0.119
Binary p	—	3.52×10^{-13}	2.47×10^{-4}

Table 2: Regression Results for Sampling-Induced Explanation Instability

Table II shows that the results of the regressions were the same for all three datasets. Univariate r and p show the bivariate correlation between the exclusion fraction and SOFIA-Index, and Multi. Multivariate coefficient and significance are displayed in Beta and Multi. p. Partial R2 is the R2 of the model that excludes the explained variance of exclusion fraction of the SOFIA-Index; Binary Beta and Binary p are the binary top-k exclusion indicator. In general the results corroborate the study hypotheses in all datasets. The correlation between exclusion fraction and SOFIA-Index on Cora is insignificant, and ACTOR and Reddit have significant and positive univariate and multivariate relationships. The largest values are for Actor (Partial $R^2 = 0.130$) and Reddit (Partial $R^2 = 0.095$), which indicates that exclusion fraction is a good predictor of the instability of explanation, particularly for heterophilous graphs.

4.6 Node-Level Instability Scatter Analysis

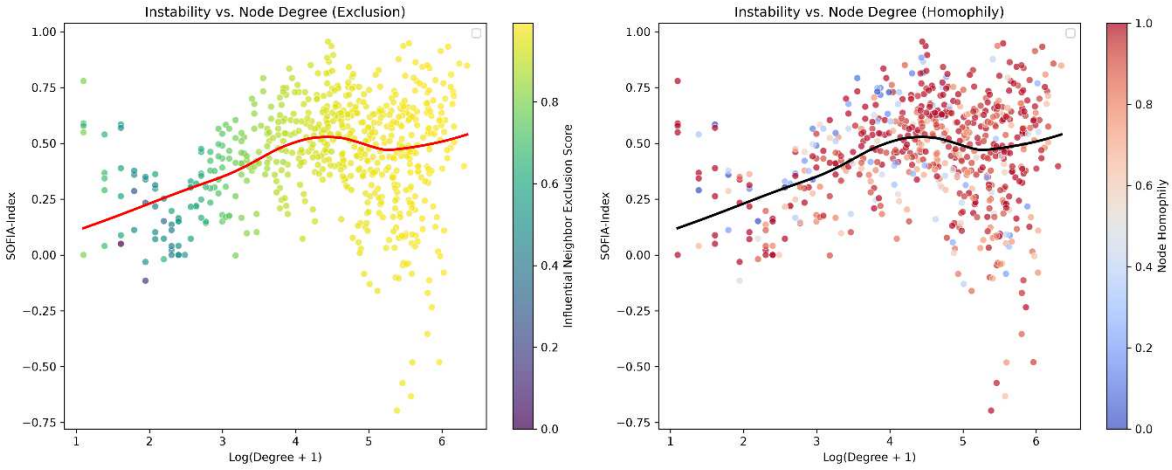


FIGURE 12: Instability vs. Node Degree — Reddit

Figure 12 presents the relationship between log-scaled node degree and SOFIA-Index for Reddit, with nodes colored by influential exclusion score in the left panel and by local homophily in the right panel. The positive regression trend in the left panel confirms that SOFIA-Index increases with node degree on this high-degree graph, consistent with the mechanism that higher-degree nodes suffer more absolute neighbor exclusion under fixed fan-out sampling. The color gradient in the left panel reveals a critical pattern: nodes with high SOFIA-Index values are disproportionately yellow (high influential exclusion scores), while nodes with low exclusion scores (purple) cluster near zero instability regardless of their degree level. This confirms that it is specifically the exclusion of perturbation-identified influential neighbors, not degree alone, that drives explanation drift — a finding that distinguishes Sofia's influence-aware analysis from a simpler degree-based account. The right panel, colored by local homophily, shows predominantly warm red coloring throughout, indicating that Reddit's high homophily does not mitigate the instability effect when exclusion is sufficiently aggressive, and that the degree-driven exclusion mechanism dominates on this graph.

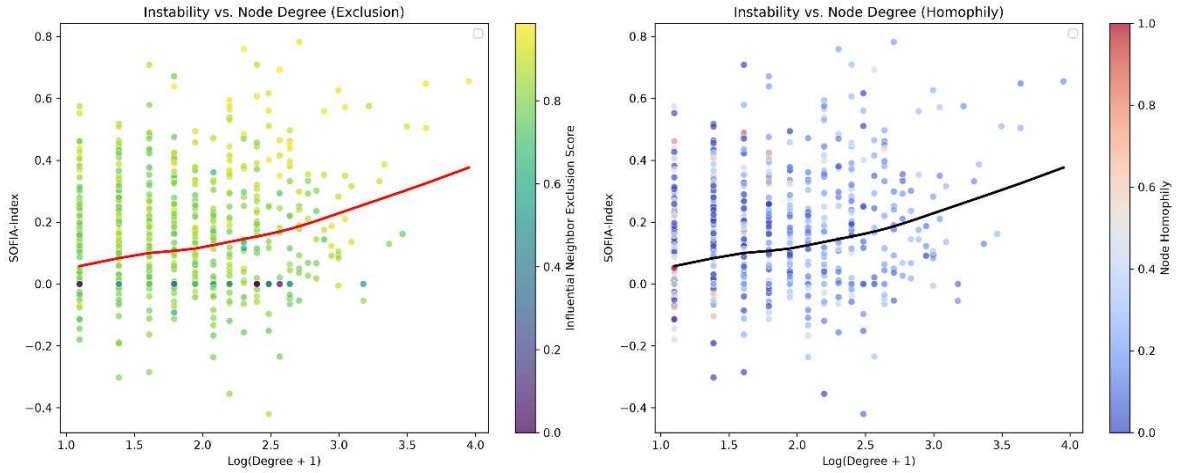


FIGURE 13: Instability vs. Node Degree — Actor

Figure 13 presents the equivalent scatter analysis for Actor. The left panel shows a positive regression trend similar to Reddit but with considerably more spread at lower degree values, reflecting Actor's more heterogeneous per-node exclusion patterns. Crucially, the color distribution in the left panel on Actor differs from Reddit: the relationship between exclusion score color and SOFIA-Index position is less cleanly stratified, reflecting that on a heterophilous graph the mechanism through which exclusion translates into instability is more complex than on a homophilous graph — a given excluded neighbor may carry far more or far less signal than average depending on whether it is one of the few same-class neighbors in a heterophilous local neighborhood. The right panel, colored by local homophily, shows that nodes with low local homophily (blue, indicating heterophilous neighborhoods) are distributed throughout the instability range, while nodes with high local homophily (red) appear more concentrated in the moderate-to-high SOFIA-Index range, providing qualitative support for the signal-concentration mechanism underlying H2.

4.7 Node-Level Qualitative Evidence

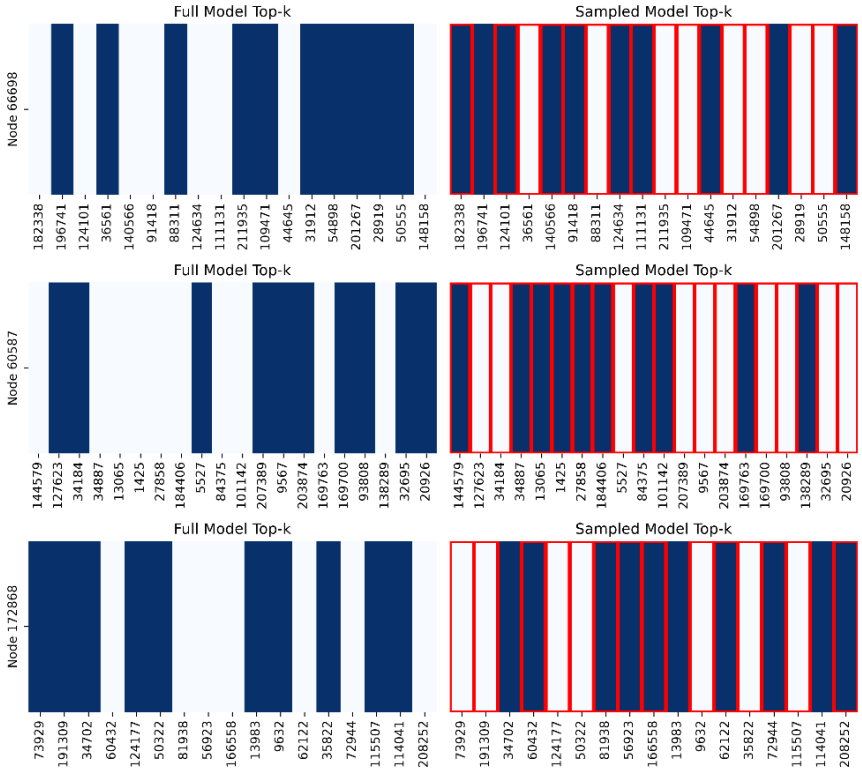


FIGURE 14: Explanation heatmap samples — Reddit

Figure 14 provides qualitative node-level evidence of sampling-induced explanation drift on Reddit for three representative nodes (66698, 60587, and 172868). For each node, the left panel shows the top-k neighbors identified as important by the full-neighborhood model and the right panel shows the top-k identified by the sampled-training model. Neighbors highlighted with red outlines in the right panel indicate positions where the sampled model selected a different neighbor than the full model. Across all three example nodes, the sampled model produces markedly different top-k neighbor sets, with the majority of positions replaced by different neighbors relative to the full model's output. This qualitative pattern is consistent with the mean cross-model Jaccard of 0.2920 reported in Table 1, and concretely illustrates how a practitioner relying on the sampled model's explanation would receive a fundamentally different account of which neighbors drove the prediction relative to the account produced by a model trained without sampling constraints. The

red-outlined positions in Figure 14 represent precisely the neighbors that the full model identified as important but that the sampled model never reliably encountered during training, instantiating the mechanism through which the SOFIA-Index gap arises at the individual node level.

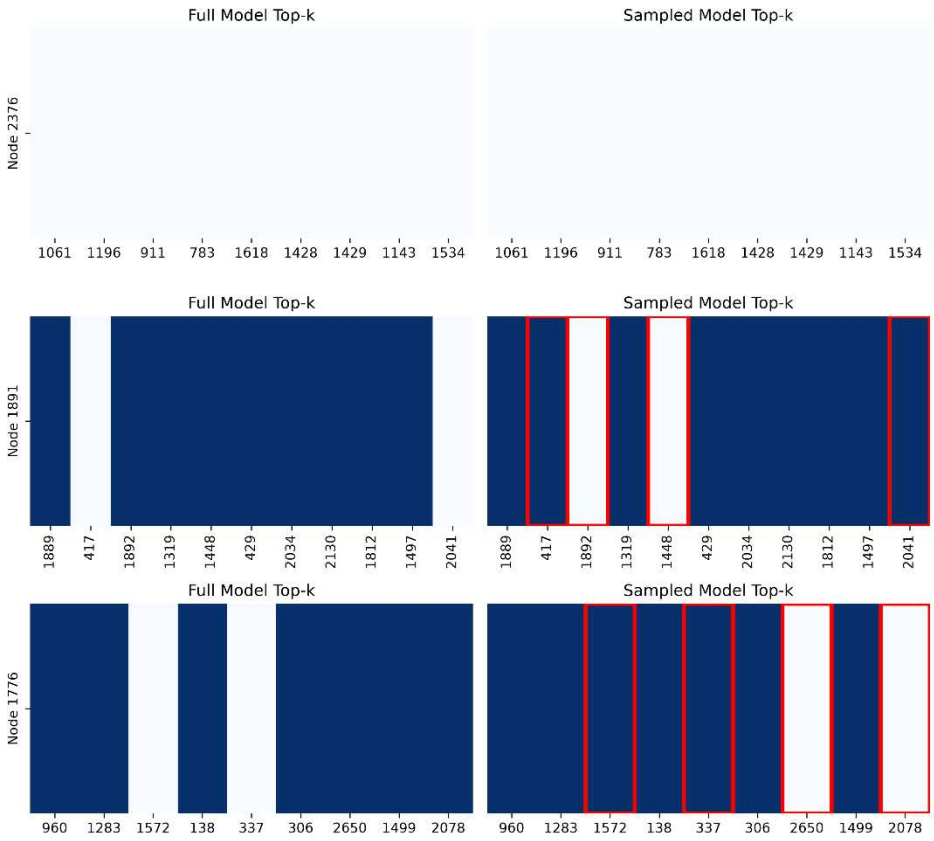


FIGURE 15: Explanation heatmap samples — Cora

Figure 15 presents the equivalent qualitative analysis for Cora. In contrast to Reddit, the top-k neighbor sets produced by the full-neighborhood and sampled models are largely consistent across all three example nodes, with few or no red-outlined differences between the left and right panels. This visual consistency directly reflects the near-zero SOFIA-Index (0.089) and non-significant regression result reported for Cora in Tables 1 and 2, and confirms that on a low-degree graph where sampling excludes few neighbors, explanation outputs remain stable across training conditions. It is the qualitative comparison between the two figures, with the same architecture but different kinds of graphs, that can be considered as the best evidence for the main thesis of this

work, namely that the stability of the explanation differs greatly, depending on the type of the graph, but otherwise keeping all other things the same, both the architecture and the procedure used by the explainer and the evaluator.

5 Discussion

The results establish, for the first time, that training-time neighbor sampling introduces systematic, measurable instability in post-hoc explanation outputs that exceeds ordinary explainer stochasticity, is significantly predicted by influential-neighbor exclusion during training, and varies in magnitude and mechanism with graph structure. The contrast between Cora (null effect, low exclusion, high homophily), Reddit (strong effect, high exclusion, high homophily), and Actor (moderate but significant effect, moderate exclusion, low homophily) maps onto the two predicted mechanisms: a degree-driven exclusion mechanism on Reddit, and a signal-concentration mechanism on Actor where heterophilous structure makes influential neighbors less substitutable, producing proportionally larger drift at lower exclusion, reflected in the higher partial R^2 (0.130 vs 0.095). Sampled models matched full-neighborhood accuracy throughout, ruling out model degradation and leaving training-time visibility as the causal variable. They have several limitations, such as the experiments using a single GCN backbone and two explainers from the same perturbation-based family, generalization across architectures and explainer families is not covered, sampling aggressiveness is binary instead of continuously swept, only node classification is examined, and the binary Top-10 indicator becomes degenerate on high-degree graphs where it should be continuous.There are several limitations: only the node classification is examined, influential neighbors are identified based on the full-neighborhood model and not on the ground-truth measure of structural importance, a continuous version of the Top-10 indicator would be more robust, the experiments use a single GCN backbone and two explainers from the same perturbation-based family, and generalization across architectures and explainer families is not covered. The main conclusion, that a seemingly innocuous engineering choice at training time will subtly influence the post-hoc explanations produced, sparks a need for more careful interpretation of GNN explanations in deployed systems and offers motivation for the development of sampling-aware correction mechanisms for the explainability pipelines.

6 Conclusion

This paper introduced Sofia, a framework for quantifying sampling-induced explanation instability in Graph Neural Networks, and presented the first systematic study of how training-time neighbor sampling affects post-hoc explainer outputs. By training matched model pairs differing only in training-time neighborhood visibility and evaluating agreement against an empirically measured per-node noise floor with degree control, Sofia isolates a previously unexamined pathway: that post-hoc explanations may reflect not a node's structural role but the accident of which neighbors a sampler showed the model during training. This pathway proves real, measurable, and structure-dependent — significant on Reddit ($\beta = 0.687$, $p = 3.33\times10^{-15}$) and Actor ($\beta = 0.255$, $p < 0.001$), consistent with heterophily amplifying instability through signal concentration, yet absent on Cora ($\beta = -0.116$, $p = 0.401$) — while sampled models matched full-neighborhood accuracy throughout, leaving training-time visibility as the causal variable. Three methodological contributions stand independently: per-node empirical noise floor derivation, influence-aware exclusion measurement, and a cross-dataset design distinguishing degree-driven exclusion from signal concentration. Limitations include a single GCN backbone, perturbation-based explainers only, binary fan-out treatment, and restriction to node classification. For practitioners, the SOFIA-Index offers a lightweight per-node diagnostic for sampling-contaminated explanations; the full implementation is available at https://github.com/PRABU-595/sofia. Future work should extend evaluation across architectures and explainer families, sweep fan-out continuously, test generalization to link prediction and graph-level tasks, and investigate sampling-aware correction mechanisms — a trade-off between computational efficiency and explanation reliability that this work most strongly recommends pursuing.

References

1. Kipf, T. N., & Welling, M. (2017). Semi-Supervised Classification with Graph Convolutional Networks. International Conference on Learning Representations (ICLR).
2. Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y. (2017). Graph attention networks. *stat*, *1050*(20), 10-48550.
3. Xu, K., Hu, W., Leskovec, J., & Jegelka, S. (2019). How Powerful are Graph Neural

Networks? International Conference on Learning Representations (ICLR).

4. Hamilton, W. L., Ying, Z., & Leskovec, J. (2017). Inductive Representation Learning on Large Graphs. Advances in Neural Information Processing Systems (NeurIPS), 30, 1025–1035.
5. Chen, J., Ma, T., & Xiao, C. (2018). Fastgcn: fast learning with graph convolutional networks via importance sampling. *arXiv preprint arXiv:1801.10247*.
6. Chiang, W. L., Liu, X., Si, S., Li, Y., Bengio, S., & Hsieh, C. J. (2019, July). Cluster-gcn: An efficient algorithm for training deep and large graph convolutional networks. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 257-266).
7. Zeng, H., Zhou, H., Srivastava, A., Kannan, R., & Prasanna, V. (2019). Graphsaint: Graph sampling based inductive learning method. *arXiv preprint arXiv:1907.04931*.
8. Zou, D., Hu, Z., Wang, Y., Jiang, S., Sun, Y., & Gu, Q. (2019). Layer-Dependent Importance Sampling for Training Deep and Large Graph Convolutional Networks. Advances in Neural Information Processing Systems (NeurIPS), 32.
9. Wu, F., Souza, A., Zhang, T., Fifty, C., Yu, T., & Weinberger, K. (2019, May). Simplifying graph convolutional networks. In *International conference on machine learning* (pp. 6861-6871). Pmlr.
10. Younesian, T., Thanapalasingam, T., van Krieken, E., Daza, D., & Bloem, P. (2023, October). Grapes: Learning to sample graphs for scalable graph neural networks. In *NeurIPS 2023 Workshop: New Frontiers in Graph Learning*.
11. Balın, M. F., & Çatalyürek, Ü. V. (2022). Layer-Neighbor Sampling--Defusing Neighborhood Explosion in GNNs. *arXiv preprint arXiv:2210.13339*.
12. Zhang, B., Kuppannagari, S. R., Kannan, R., & Prasanna, V. (2021, September). Efficient neighbor-sampling-based gnn training on cpu-fpga heterogeneous platform. In *2021 IEEE High Performance Extreme Computing Conference (HPEC)* (pp. 1-7). IEEE.
13. Min, S. W., Wu, K., Hidayetoglu, M., Xiong, J., Song, X., & Hwu, W.-m. (2022). Graph Neural Network Training and Data Tiering. Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 3555–3565. https://doi.org/10.1145/3534678.3539038

14. Gandhi, S., & Iyer, A. P. (2021). P3: Distributed deep graph learning at scale. In *15th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 21)* (pp. 551-568).
15. Ying, Z., Bourgeois, D., You, J., Zitnik, M., & Leskovec, J. (2019). GNNExplainer: Generating Explanations for Graph Neural Networks. Advances in Neural Information Processing Systems (NeurIPS), 32.
16. Luo, D., Cheng, W., Xu, D., Yu, W., Zong, B., Chen, H., & Zhang, X. (2020). Parameterized explainer for graph neural network. *Advances in neural information processing systems*, *33*, 19620-19631.
17. Vu, M., & Thai, M. T. (2020). Pgm-explainer: Probabilistic graphical model explanations for graph neural networks. *Advances in neural information processing systems*, *33*, 12225-12235.
18. Schlichtkrull, M. S., De Cao, N., & Titov, I. (2020). Interpreting graph neural networks for NLP with differentiable edge masking. *arXiv preprint arXiv:2010.00577*.
19. Yuan, H., Tang, J., Hu, X., & Ji, S. (2020, August). Xgnn: Towards model-level explanations of graph neural networks. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 430-438).
20. Yuan, H., Yu, H., Wang, J., Li, K., & Ji, S. (2021, July). On explainability of graph neural networks via subgraph explorations. In *International conference on machine learning* (pp. 12241-12252). PMLR.
21. Zhang, S., Liu, Y., Shah, N., & Sun, Y. (2022). Gstarx: Explaining graph neural networks with structure-aware cooperative games. *Advances in neural information processing systems*, *35*, 19810-19823.
22. Wang, X., Wu, Y., Zhang, A., He, X., & Chua, T. S. (2021). Towards multi-grained explainability for graph neural networks. *Advances in neural information processing systems*, *34*, 18446-18458.
23. Yuan, H., Yu, H., Gui, S., & Ji, S. (2023). Explainability in Graph Neural Networks: A Taxonomic Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(5), 5782–5799. https://doi.org/10.1109/TPAMI.2022.3204236
24. Schnake, T., Eberle, O., Lederer, J., Nakajima, S., Schütt, K. T., Müller, K.-R., & Montavon, G. (2022). Higher-Order Explanations of Graph Neural Networks via Relevant Walks. IEEE

Transactions on Pattern Analysis and Machine Intelligence, 44(11), 7581–7596. https://doi.org/10.1109/TPAMI.2021.3115452

25. Wei, F., & Mei, K. (2024). Towards Self-Explainable Graph Convolutional Neural Network with Frequency Adaptive Inception. Pattern Recognition, 146, 109991. https://doi.org/10.1016/j.patcog.2023.109991
26. Agarwal, C., Queen, O., Lakkaraju, H., & Zitnik, M. (2023). Evaluating Explainability for Graph Neural Networks. Scientific Data, 10(1), 144.
27. Agarwal, C., Zitnik, M., & Lakkaraju, H. (2022). Probing GNN Explainers: A Rigorous Theoretical and Empirical Analysis of GNN Explanation Methods. Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS), PMLR 151, 8969–8996.
28. Agarwal, C., Queen, O., Lakkaraju, H., & Zitnik, M. (2022). An Explainable AI Library for Benchmarking Graph Explainers. ACM Web Conference Workshop on Graph Learning Benchmarks (GLB).
29. Amara, K., Ying, Z., Zhang, Z., Han, Z., Zhao, Y., Shan, Y., Brandes, U., Schemm, S., & Zhang, C. (2022). GraphFramEx: Towards Systematic Evaluation of Explainability Methods for Graph Neural Networks. Learning on Graphs Conference (LoG), PMLR.
30. Faber, L., Moghaddam, A. K., & Wattenhofer, R. (2021). When Comparing to Ground Truth is Wrong: On Evaluating GNN Explanation Methods. Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 332–341.
31. Agarwal, C., Krishna, S., Saxena, E., Pawelczyk, M., Johnson, N., Puri, I., ... & Lakkaraju, H. (2022). Openxai: Towards a transparent evaluation of model explanations. *Advances in neural information processing systems*, *35*, 15784-15799.
32. Kosan, M., Verma, S., Armgaan, B., Pahwa, K., Singh, A., Medya, S., & Ranu, S. (2023). Gnnx-bench: Unravelling the utility of perturbation-based gnn explainers through in-depth benchmarking. *arXiv preprint arXiv:2310.01794*.
33. Zhu, J., Yan, Y., Zhao, L., Heimann, M., Akoglu, L., & Koutra, D. (2020). Beyond homophily in graph neural networks: Current limitations and effective designs. *Advances in neural information processing systems*, *33*, 7793-7804.
34. Zhu, J., Rossi, R. A., Rao, A., Mai, T., Lipka, N., Ahmed, N. K., & Koutra, D. (2021, May).

Graph neural networks with heterophily. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 35, No. 12, pp. 11168-11176).

35. Luan, S., Hua, C., Lu, Q., Zhu, J., Zhao, M., Zhang, S., ... & Precup, D. (2022). Revisiting heterophily for graph neural networks. *Advances in neural information processing systems*, *35*, 1362-1375.
36. Pei, H., Wei, B., Chang, K. C. C., Lei, Y., & Yang, B. (2020). Geom-gcn: Geometric graph convolutional networks. *arXiv preprint arXiv:2002.05287*.
37. Platonov, O., Kuznedelev, D., Diskin, M., Babenko, A., & Prokhorenkova, L. (2023). A critical look at the evaluation of GNNs under heterophily: Are we really making progress?. *arXiv preprint arXiv:2302.11640*.
38. Lim, D., Hohne, F., Li, X., Huang, S. L., Gupta, V., Bhalerao, O., & Lim, S. N. (2021). Large scale learning on non-homophilous graphs: New benchmarks and strong simple methods. *arXiv preprint arXiv:2110.14446*.
39. Sen, P., Namata, G., Bilgic, M., Getoor, L., Galligher, B., & Eliassi-Rad, T. (2008). Collective classification in network data. *AI magazine*, *29*(3), 93-93.
40. Hu, W., Fey, M., Zitnik, M., Dong, Y., Ren, H., Liu, B., ... & Leskovec, J. (2020). Open graph benchmark: Datasets for machine learning on graphs. *Advances in neural information processing systems*, *33*, 22118-22133.
41. Rozemberczki, B., Allen, C., & Sarkar, R. (2021). Multi-Scale Attributed Node Embedding. Journal of Complex Networks, 9(2).
42. Fey, M., & Lenssen, J. E. (2019). Fast graph representation learning with PyTorch Geometric. *arXiv preprint arXiv:1903.02428*.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SF.pdf