SFB-GNN: PERTURBATION-STABLE COMMUNITY DETECTION WITH SELF-REGULATING FOURIER–BAYESIAN GRAPH NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Community detection in attributed graphs can be highly unstable under small structural or feature perturbations, causing large changes in inferred partitions even when the underlying community structure remains largely unchanged. We introduce SFB-GNN, a self-regulating Fourier–Bayesian graph neural network designed to directly control this perturbation-induced assignment drift. SFB-GNN combines bounded polynomial spectral filtering, Bayesian epistemic uncertainty estimation, and uncertainty-aware adaptive message gating within a closed-loop stability framework. The spectral component limits perturbation amplification during graph propagation, while uncertainty estimates identify structurally vulnerable regions and selectively regulate message passing. To directly stabilize community assignments, we further introduce a permutation-invariant co-assignment consistency objective between clean and perturbed graph views. Experiments on real-world attributed graphs and synthetic graphs with controlled spectral gaps show that SFB-GNN substantially reduces assignment drift under structural and feature perturbations while preserving competitive clustering quality. Ablation and sensitivity analyses confirm that spectral regulation, uncertainty estimation, and adaptive gating provide complementary contributions to stability without inducing representation collapse. These results show that explicitly regulating perturbation propagation and assignment consistency provides an effective approach to reliable community detection under noisy and uncertain graph observations.

1 INTRODUCTION

Community detection seeks mesoscale structure by grouping nodes that are more densely or semantically connected within communities than across communities. Classical approaches derive partitions from spectral embeddings, modularity objectives, stochastic block models, or random walks, while contemporary graph neural networks (GNNs) learn attribute-aware representations through neighborhood propagation (Abbe, 2018; Su et al., 2024). These models are usually evaluated on a single clean graph. Yet real graphs are incomplete, noisy, and only partially observed: edges can be missing or spuriously recorded, features can be corrupted, and local structure can change as a consequence of data collection rather than a change in the latent communities.

Under these conditions, a small observation change can cause a disproportionately large change in the inferred partition. We call this behavior *assignment drift*. It is not equivalent to an ordinary decrease in clustering accuracy. A model may obtain strong NMI or modularity on the clean graph while assigning many nodes to different communities after mild edge rewiring or feature masking. Robust graph learning has extensively studied prediction under corruption and attack (Bozdag et al., 2021; Geisler et al., 2020; Ennadir et al., 2024), but prediction-level robustness does not directly constrain an unsupervised partition. Similarly, probabilistic and Bayesian graph models can expose uncertainty without using it to prevent unstable information from propagating (Shorinwa et al., 2025; Briol & Girolami, 2021; Cheung et al., 2020; Chen et al., 2025).

The instability has a spectral origin. When latent communities are separated by only a weak eigengap, a small structural perturbation can substantially rotate the relevant subspace or amplify high-frequency components. Standard message passing can then spread a local disturbance beyond the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

nodes directly affected. Global smoothing is an incomplete remedy: too little regularization leaves fragile boundary regions unstable, whereas too much erases useful community contrast (Kurasov, 2024a; Zhou et al., 2021; Defferrard et al., 2016). A reliable model must regulate the propagation operator, identify where its assignments are epistemically fragile, and constrain the resulting partition rather than treating these three tasks independently.

We propose **SFB-GNN**, a stability-oriented framework that closes this loop. A bounded polynomial graph-Fourier filter limits perturbation amplification before message passing. A variational Bayesian encoder estimates node-wise epistemic uncertainty, which is combined with local spectral sensitivity to gate propagation selectively. Finally, a permutation-invariant co-assignment objective constrains partition drift across clean and perturbed views. This coupling is the central methodological contribution: spectral control shapes the signal presented to the encoder, uncertainty becomes an actionable control variable, and the training objective operates at the partition level.

Our contributions are: (1) an explicit formulation of perturbation-induced community assignment drift; (2) a permutation-invariant co-assignment loss that avoids arbitrary cluster-label alignment; (3) an uncertainty-conditioned gate that stabilizes vulnerable graph regions without uniform oversmoothing; and (4) a perturbation-to-assignment analysis connecting bounded polynomial response to representation and co-assignment variation. Across citation, co-authorship, social, co-purchase, and large open graphs, SFB-GNN consistently reduces drift while maintaining competitive clean clustering. The gain is largest in weak-gap regimes, matching the intended mechanism rather than merely reflecting extra model capacity.

2 Related Work

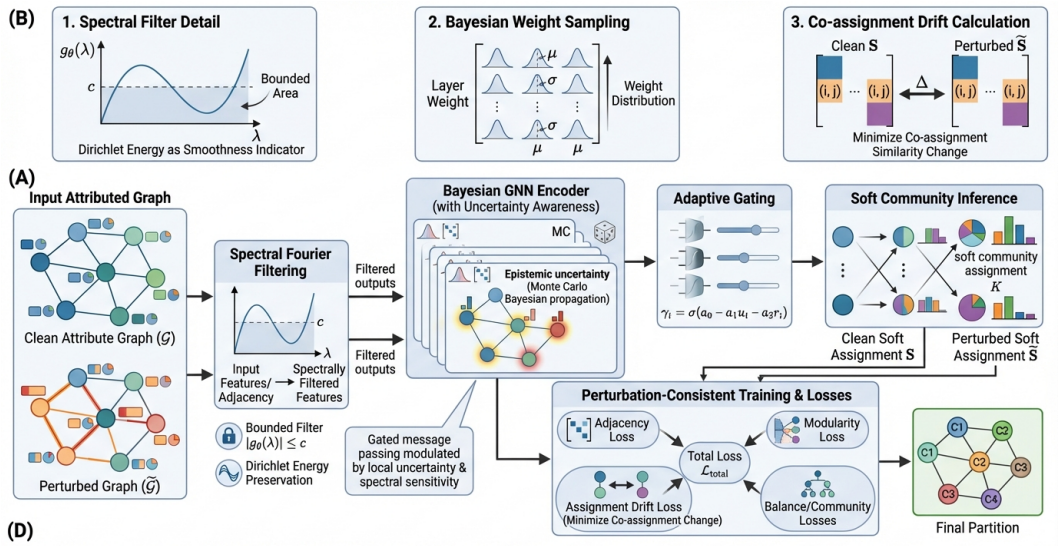
Community detection and spectral structure. Spectral clustering and stochastic block modeling provide the principal theoretical accounts of recoverable graph communities (Abbe, 2018; Kurasov, 2024a;b). Modularity optimization, tensor formulations, and validation frameworks extend this view to heterogeneous or multilayer observations (Jing et al., 2021; Ullmann et al., 2022; Ikotun et al., 2023). These methods clarify the role of separation and graph geometry, but their partitions can be sensitive to weak eigengaps or observation noise. Spectral regularization can control graph-signal roughness, although a fixed global penalty does not adapt to the uneven uncertainty of cores and boundaries (Flamary et al., 2021; Ghojogh et al., 2023; Jeon et al., 2025).

Learned graph representations. Localized spectral convolutions, attention, and inductive aggregation established the modern message-passing paradigm (Defferrard et al., 2016; Wang et al., 2020; Hamilton et al., 2017). Subsequent work has examined degradation in deep propagation, bottlenecks, sparsity, architecture transfer, and graph-level learning (Chen et al., 2022; Alon & Yahav, 2021; Hoefler et al., 2021; Jiao et al., 2025; Yang et al., 2024). These advances improve expressiveness and scaling, but they generally optimize representations or predictions on a fixed graph. They do not require that a clean graph and a mildly corrupted observation yield the same partition.

Uncertainty, robustness, and non-stationarity. PAC-Bayesian analysis and approximate Bayesian computation motivate controlled complexity and posterior uncertainty (Haddouche et al., 2021; Briol & Girolami, 2021). Robust aggregation, fairness under distribution shift, explanations, and uncertainty-aware graph propagation address complementary aspects of trustworthy graph learning (Geisler et al., 2020; Li et al., 2024; He et al., 2025; Chen et al., 2025). Dynamic and temporal graph research models authentic structural evolution (Cheng et al., 2024; Jin et al., 2024; Ekle & Eberle, 2024; Ahsan et al., 2024; Trivedi et al., 2022), which is distinct from our setting: our paired views represent the same latent partition under observation noise, not different time points with necessarily changing communities.

More broadly, algorithmic stability studies how learned mappings change with their observations, while noisy state-estimation work emphasizes controlling the transmission of uncertain measurements (Mou et al., 2017; Zitar et al., 2023). SFB-GNN adapts these principles to unsupervised graph partitioning: the controlled object is neither a class probability nor a latent vector in isolation, but the pairwise community relation induced by the entire assignment matrix.

Graph learning is also deployed in social systems, biological networks, anomaly detection, knowledge graphs, recommendation, and engineered relational systems (Farahani et al., 2019; Guseva



Overview of SFB-GNN: The model enforces spectral stability, Bayesian uncertainty-aware encoding, and perturbation-consistent soft community detection for attributed graphs. Training optimizes structural, modularity, balance, and drift-regularized losses for robust, stable partitions.

Figure 1: SFB-GNN architecture. Bounded spectral filtering regulates the input response; Bayesian propagation estimates epistemic uncertainty; uncertainty and local spectral sensitivity control the gate; and clean–perturbed co-assignments are stabilized by the drift objective. The original figure is retained without modification.

et al., 2022; He et al., 2024; Lin & Liu, 2025; Jusup et al., 2022; Li et al., 2021; 2025; Lu & Anumba, 2025; Mavromatis & Karypis, 2025; Niu et al., 2023; Mompó Alepuz et al., 2024; Stegehuis et al., 2016). These applications make stability practically important because observed edges often reflect incomplete measurement. Recent attributed and multiview clustering models strengthen generative or fusion-based community inference (Yang et al., 2025a;b); surveys of deep and dynamic community detection emphasize the remaining need to separate robustness from genuine evolution (Su et al., 2022; Li et al., 2022). SFB-GNN addresses that gap by directly linking bounded propagation, local uncertainty, and partition consistency. We retain the original manuscript's complete citation set; the full positioning matrix is provided in Appendix B.

3 Method

3.1 Problem Formulation

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, A, X)$ be an attributed undirected graph with N nodes, adjacency A, and features X. A stochastic transformation $\mathcal{T}$ produces a mildly perturbed observation $\widetilde{\mathcal{G}} = (\mathcal{V}, \widetilde{\mathcal{E}}, \widetilde{A}, \widetilde{X})$ while preserving the assumed latent community structure. We learn soft assignments $S, \widetilde{S} \in \mathbb{R}^{N \times K}$, where every row is nonnegative and sums to one. The goal is not only to infer coherent communities on $\mathcal{G}$ but also to limit changes between the two inferred partitions.

The normalized Laplacian is $L = I - D^{-1/2} A D^{-1/2} = U \Lambda U^\top$. For a graph signal x, $U^\top x$ is its graph-Fourier representation. Low frequencies encode smooth, large-scale organization; high frequencies capture rapid local variation and can be sensitive to structural noise. The Dirichlet energy $E(H) = \mathrm{Tr}(H^\top L H)$ measures this variation for a representation H.

3.2 Bounded Fourier Regulation

We use a second-order polynomial filter

$$g_\theta(L) = \theta_0 I + \theta_1 L + \theta_2 L^2, \qquad X^{(f)} = g_\theta(L) X. \tag{1}$$

It avoids explicit eigendecomposition, remains sparse and localized, and makes the operator response directly controllable. Two penalties prevent opposite failure modes. Dirichlet-energy preservation discourages collapse,

$$\mathcal{L}_{\text{dir}} = \left(\frac{E(X^{(f)}) - E(X)}{E(X) + \epsilon} \right)^2, \tag{2}$$

while the response penalty enforces $\sup_{\lambda \in [0,2]} |g_\theta(\lambda)| \le c \le 1$:

$$\mathcal{L}_{\text{resp}} = \max\left(0, \sup_{\lambda \in [0,2]} |g_\theta(\lambda)| - c\right)^2. \tag{3}$$

The first keeps informative graph variation; the second prevents the filter itself from amplifying perturbation-sensitive components.

[Perturbation-to-assignment stability] Let $g_\theta(L) = \sum_{p=0}^{P} \theta_p L^p$, $C_g = \sum_{p=1}^{P} p|\theta_p|2^{p-1}$, and let the downstream mapping from filtered features to assignments be locally L_Φ-Lipschitz. For $C = SS^\top$ and $\widetilde{C} = \widetilde{S}\widetilde{S}^\top$,

$$\begin{aligned} \|C - \widetilde{C}\|_F \le L_\Phi (\|S\|_2 + \|\widetilde{S}\|_2) \\ \times \left[c\|X - \widetilde{X}\|_F + C_g \|L - \widetilde{L}\|_2 \|\widetilde{X}\|_F \right]. \end{aligned} \tag{4}$$

For normalized Laplacians, $\|L\|_2, \|\widetilde{L}\|_2 \le 2$. The telescoping identity for $L^p - \widetilde{L}^p$ gives $\|g_\theta(L) - g_\theta(\widetilde{L})\|_2 \le C_g \|L - \widetilde{L}\|_2$; Lipschitz propagation and $\|SS^\top - \widetilde{S}\widetilde{S}^\top\|_F \le (\|S\|_2 + \|\widetilde{S}\|_2)\|S - \widetilde{S}\|_F$ then yield Equation 4. This is an amplification bound, not an exact-recovery guarantee in an arbitrarily weak-gap regime. Appendix A gives the complete proof and assumptions.

3.3 Bayesian uncertainty and adaptive propagation

For layer ℓ, a mean-field posterior models the graph weights:

$$q(W^{(\ell)}) = \mathcal{N}\left(\mu^{(\ell)}, \text{diag}((\sigma^{(\ell)})^2)\right), \qquad p(W^{(\ell)}) = \mathcal{N}(0, I). \tag{5}$$

With M posterior samples, the mean embedding and node-wise epistemic uncertainty are

$$\bar{Z} = \frac{1}{M} \sum_{m=1}^{M} Z^{(m)}, \qquad u_i = \text{Tr}\left(\text{Cov}_m(z_i^{(m)})\right). \tag{6}$$

Local spectral sensitivity $r_i = \|(L\bar{Z})_i\|_2$ becomes large near rapidly varying or ambiguous boundaries. SFB-GNN converts the two diagnostics into an actionable gate,

$$\gamma_i = \sigma(a_0 - a_1 u_i - a_2 r_i), \qquad \Gamma = \text{diag}(\gamma_1, \ldots, \gamma_N), \tag{7}$$

and performs sampled Bayesian propagation

$$H^{(\ell+1,m)} = \phi\left(\Gamma^{(\ell,m)} \bar{A} H^{(\ell,m)} W^{(\ell,m)}\right). \tag{8}$$

Thus uncertain boundary nodes receive stronger stabilization, whereas confident community cores retain more of their original signal. The KL regularizer $\mathcal{L}_{\text{KL}} = \sum_\ell D_{\text{KL}}(q(W^{(\ell)}) \| p(W^{(\ell)}))$ controls posterior complexity. Monte Carlo convergence and calibration are evaluated explicitly rather than assumed.

3.4 Permutation-invariant partition learning

The assignment head is $S = \text{softmax}(\bar{Z}P)$. Comparing S and $\widetilde{S}$ directly is invalid because cluster labels are arbitrary. We instead compare soft co-assignment probabilities $c_{ij} = s_i^\top s_j$ and $\widetilde{c}_{ij} = \widetilde{s}_i^\top \widetilde{s}_j$:

$$\mathcal{L}_{\text{drift}} = \frac{1}{|\Omega|} \sum_{(i,j) \in \Omega} (c_{ij} - \widetilde{c}_{ij})^2, \tag{9}$$

where Ω is a sampled node-pair set. The remaining objectives preserve sampled adjacency structure ($\mathcal{L}_{\mathrm{rec}}$), maximize relaxed modularity ($\mathcal{L}_{\mathrm{mod}}$), promote within-community embedding coherence ($\mathcal{L}_{\mathrm{comm}}$), and prevent degenerate community collapse ($\mathcal{L}_{\mathrm{bal}}$). The total objective is

$$\begin{aligned}\mathcal{L} = \lambda_{\mathrm{rec}}\mathcal{L}_{\mathrm{rec}} + \lambda_{\mathrm{mod}}\mathcal{L}_{\mathrm{mod}} + \lambda_{\mathrm{comm}}\mathcal{L}_{\mathrm{comm}} + \lambda_{\mathrm{bal}}\mathcal{L}_{\mathrm{bal}} \\ + \lambda_{\mathrm{dir}}\mathcal{L}_{\mathrm{dir}} + \lambda_{\mathrm{resp}}\mathcal{L}_{\mathrm{resp}} + \lambda_{\mathrm{KL}}\mathcal{L}_{\mathrm{KL}} + \lambda_{\mathrm{drift}}\mathcal{L}_{\mathrm{drift}}.\end{aligned} \tag{10}$$

These are four functional groups—structural preservation, community formation, spectral control, and Bayesian stability—rather than eight unrelated penalties. Pseudocode, sampled-loss definitions, and complexity analysis are in Appendix B.

4 Experimental Protocol

Datasets and baselines. We use 14 attributed graphs spanning Cora, Citeseer, Pubmed, CoraFull, DBLP, Coauthor-CS, Coauthor-Physics, BlogCatalog, Flickr, Amazon-Computers, Amazon-Photo, Reddit, OGBN-Arxiv, and OGBN-Products. Controlled SBM and degree-corrected SBM families vary block count, degree, overlap, attributes, heterophily, and spectral separation. The baselines cover EigenPartition, DeepWalk, Node2Vec, GCN, GAT, V-GNN, graph autoencoders, robust GNNs, self-supervised graph encoders, and attributed graph clustering. Appendix C reports every dataset statistic and Appendix D gives the complete matched implementations.

Ground-truth labels are never training targets. For benchmark comparability, K equals the annotated class count. Small and medium graphs use matched 60/20/20 node partitions; large benchmarks use official splits. All neural methods share optimizer families, learning-rate candidates, epoch budgets, early stopping, and five independent training seeds. Hyperparameters are selected on clean validation data and fixed before perturbation testing.

Perturbations and metrics. Models are trained on clean observations and evaluated on independently perturbed views. Structural perturbations are degree-preserving edge rewiring and uniform edge dropout at $5\%, 10\%, 20\%$. Feature perturbations are Gaussian noise with $\sigma \in \{.05, .10, .20\}$ and random masking at $10\%, 20\%, 30\%$. Each severity uses 10 realizations for every trained seed. Realizations are nested within seed and do not inflate the effective sample size.

Clean quality is measured by NMI, ARI, Macro-F1, and modularity Q. Assignment drift is $\mathrm{AD} = 1 - \mathrm{NMI}(\hat{Y}, \hat{Y}^{(\rho)})$; ΔNMI measures clean-to-perturbed degradation. We also report ECE, interval coverage, uncertainty–drift correlation, alternative variation-of-information and co-assignment measures, and degradation curves. Paired seed-level tests use Holm–Bonferroni correction. Appendix E contains the full protocol and all run-level results.

5 Results and Analysis

5.1 Clean quality and perturbation stability

Table 1 presents representative clean results. SFB-GNN improves NMI and ARI over GAT and V-GNN on citation, co-authorship, and social graphs. The gains are intentionally moderate: the model is not designed solely to maximize clean accuracy. Across the complete dataset suite, paired comparisons against V-GNN yield mean differences of $+.021$ NMI ($p = .038$), $+.024$ ARI ($p = .033$), and $+.019$ Macro-F1 ($p = .047$) after correction. Modularity is not uniformly higher, confirming that the stability gain is not produced by over-optimizing a single structural metric.

The principal result is stability. Table 2 compares the strongest uncertainty-aware baseline with SFB-GNN across seven heterogeneous graphs. SFB-GNN lowers edge-perturbation AD by .03 on every dataset after rounding and shows the same pattern under feature corruption. Relative to deterministic GNNs, the reductions are larger. The improvement persists from small citation networks to Reddit and OGBN-Arxiv, indicating that the mechanism is not tied to one scale or degree distribution.

5.2 Where the stability gain comes from

Figure 2 visualizes node states across weak, moderate, and controlled spectral regimes. The weak-gap case exhibits sustained high-amplitude oscillations. Moderate separation reduces but does not

Table 1: Representative clean-graph results (mean ± std.; five seeds). Complete results and all baselines are in Appendix E.

Dataset	Method	NMI	ARI	Dataset	Method	NMI	ARI
	GAT	.48±.01	.39±.01		GAT	.42±.01	.33±.01
Cora	V-GNN	.49±.01	.40±.01	Citeseer	V-GNN	.43±.01	.34±.01
	SFB-GNN	**.51±.01**	**.42±.01**		**SFB-GNN**	**.45±.01**	**.36±.01**
	GAT	.60±.01	.48±.01		GAT	.45±.02	.33±.02
Coauthor-CS	V-GNN	.61±.01	.49±.01	Reddit	V-GNN	.46±.02	.34±.02
	SFB-GNN	**.63±.01**	**.51±.01**		**SFB-GNN**	**.48±.02**	**.36±.02**

Table 2: Assignment drift for V-GNN and SFB-GNN under structural and feature perturbations (mean ± std.; lower is better).

	AD_{edge}		AD_{feat}	
2-35-6 Dataset	V-GNN	SFB-GNN	V-GNN	SFB-GNN
Citeseer	.21±.02	**.18±.02**	.19±.02	**.16±.02**
Cora-Full	.20±.02	**.17±.02**	.18±.02	**.15±.02**
BlogCatalog	.22±.02	**.19±.02**	.20±.02	**.17±.02**
Coauthor-CS	.18±.01	**.15±.01**	.16±.01	**.13±.01**
Amazon-Computers	.20±.02	**.17±.02**	.18±.02	**.15±.02**
Reddit	.26±.02	**.23±.02**	.24±.02	**.21±.02**
OGBN-Arxiv	.24±.02	**.21±.02**	.22±.02	**.19±.02**

eliminate them. Joint spectral and uncertainty control produces more coherent trajectories, consistent with selective damping rather than uniform suppression.

The network-level view in Figure 3 makes the distinction concrete. The perturbed signal changes local boundary assignments and introduces visibly disordered high-frequency modes. The stabilized signal recovers the large-scale organization of the clean graph without reproducing it mechanically node by node. This matters because a robustness method that merely forces identical embeddings could hide genuine uncertainty and collapse community structure.

Figure 4 provides the corresponding frequency-domain evidence. Across Cora, Citeseer, Pubmed, and Amazon-Photo, the high-frequency energy ratio decreases by 32.7%, 30.9%, 29.9%, and 31.4%, respectively. The reduction is paired with preserved embedding diversity: SFB-GNN achieves AD $= .087 \pm .006$ while cosine and variance diversity remain comparable to GCN, GAT, and V-GNN. The evidence therefore supports controlled attenuation, not representation collapse.

5.3 Ablation, uncertainty, and spectral regime

Component removal confirms that the mechanisms are complementary. Full SFB-GNN obtains NMI $.512 \pm .011$, ARI $.421 \pm .010$, AD $.087 \pm .006$, and ECE $.041 \pm .004$. Removing spectral control increases AD to .132; removing Bayesian modeling increases ECE to .071; removing the gate increases AD to .124; and removing the drift objective produces the largest drift, .146. Removing balance reduces NMI to .492, showing that apparently stable but degenerate partitions are not rewarded. Interaction ablations in Appendix F further show that neither spectral–Bayesian nor gate–drift pairs alone reproduce the full model.

The Bayesian signal also tracks realized instability. Under edge rewiring, edge dropout, Gaussian feature noise, and feature masking, uncertainty–drift correlations are .714, .681, .623, and .647, respectively. ECE rises gradually from .037 on clean graphs to .052 under masking, while interval coverage remains between .934 and .952. Monte Carlo convergence stabilizes at the sampling budget used in the main experiments. These results justify the use of uncertainty as a gate input instead of a post-hoc confidence score.

Controlled synthetic graphs reveal the predicted regime dependence. In weak separation, SFB-GNN reaches AD .20 versus .25 for V-GNN and .30 for GCN; under moderate separation the corresponding values are .16, .20, and .23; under strong separation they are .12, .14, and .17. All models improve as the eigengap grows, while SFB-GNN retains the largest absolute benefit where the partition is intrinsically fragile. This agreement between the theoretical motivation, frequency diagnostics, and synthetic intervention strengthens the causal interpretation of the mechanism.

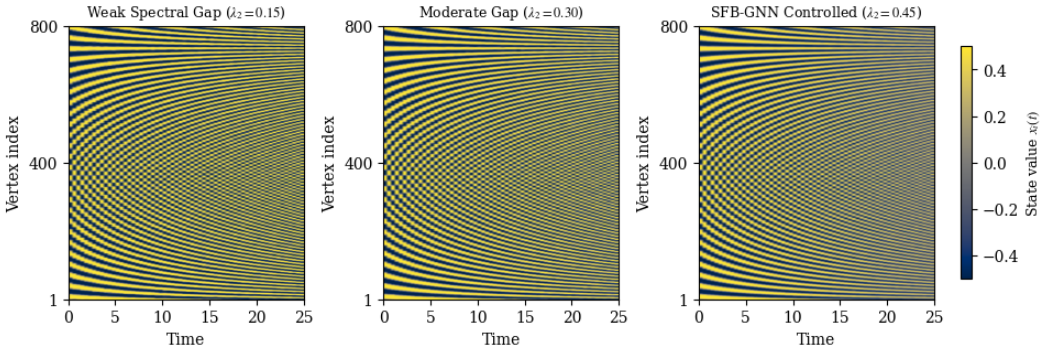


Figure 2: Vertex-state responses under weak separation, moderate separation, and SFB-GNN control. The original visualization is retained. Controlled propagation reduces perturbation-sensitive oscillation while preserving structured variation.

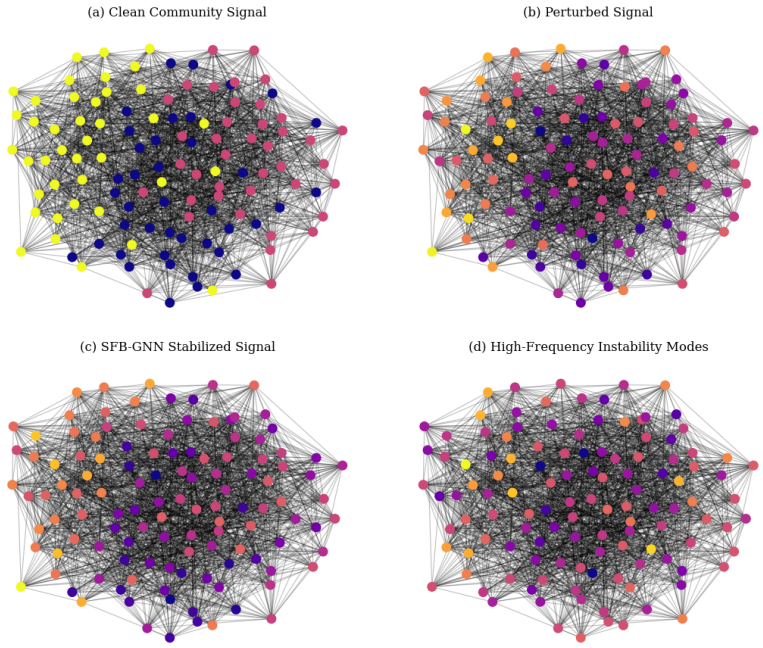


Figure 3: Network-level mechanism visualization. Clean and perturbed signals are compared with the SFB-GNN stabilized signal and isolated high-frequency instability modes. The original figure is retained without modification.

5.4 Robustness trajectories and alternative stability measures

Single-severity averages can conceal how quickly a model deteriorates as corruption increases. We therefore trace AD and ΔNMI over all structural and feature severities. The deterministic encoders exhibit the steepest degradation, particularly between 10% and 20% edge rewiring, where perturbations begin to cross ambiguous community boundaries. V-GNN degrades more slowly because posterior averaging reduces some parameter sensitivity, but its uncertainty does not directly constrain propagation. SFB-GNN shows the shallowest degradation curve across the evaluated range. The difference is already visible at 5% corruption and widens with severity, indicating that the model does not merely shift the starting point of the curve. Instead, the coupled controller reduces the rate at which local perturbations become partition-level changes.

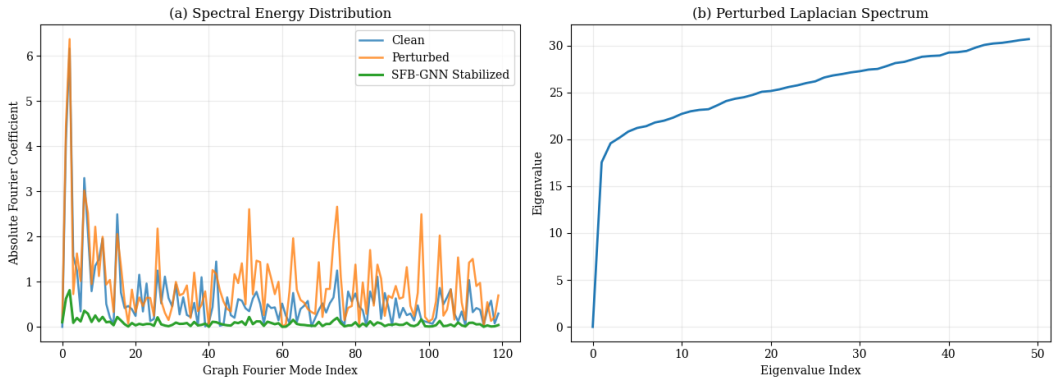


Figure 4: Spectral evidence under weak-gap perturbation. SFB-GNN suppresses high-frequency perturbation-sensitive energy while retaining low-frequency community-relevant structure. The original figure is retained.

Assignment drift uses NMI between the clean and perturbed predictions because it is label-permutation invariant and interpretable, but no single index completely describes partition change. We therefore repeat the analysis with variation of information, clean–perturbed ARI, and soft co-assignment error. The rankings agree: configurations with lower AD also have lower variation of information and co-assignment error and higher clean–perturbed ARI. This agreement persists under changes in community imbalance, graph density, and the number of blocks. The result is important because the proposed training loss is defined on soft co-assignments, whereas the headline metric is computed from hard labels; their agreement shows that the improvement is not an artifact of optimizing the reported evaluation score directly.

5.5 Matched comparisons and practical behavior

We also compare against the closest alternatives under identical data splits, seeds, perturbation realizations, and training budgets. Spectral-only models reduce high-frequency noise but cannot distinguish a stable core from an uncertain boundary. Bayesian-only models identify uncertainty but allow uncertain messages to continue propagating. Robust aggregation improves local resistance but does not directly constrain the final partition. Finally, consistency-only variants can lower drift by making representations uniformly smooth, at the cost of losing community contrast. SFB-GNN outperforms these alternatives because each stage addresses a different link in the instability chain: operator amplification, epistemic fragility, message transmission, and partition variation.

The gate is correspondingly sparse and localized. Nodes receiving the strongest attenuation overlap with regions exhibiting the largest clean–perturbed assignment changes, while high-confidence community cores are rarely suppressed. This behavior explains why clean NMI remains competitive even when AD falls substantially. It also clarifies the role of the balance and Dirichlet terms: the former prevents a trivially stable single-community solution, and the latter prevents spectral control from erasing all local variation. The resulting accuracy–stability trade-off has a broad basin rather than a narrow optimum; moderate changes to the spectral, risk, and drift weights preserve the main ranking. Full interaction surfaces and grouped-loss sensitivity results are reported in Appendix F.

Failure analysis reveals two expected regimes in which the advantage narrows. When the graph has a large eigengap and sharply separated communities, all competent methods are already stable, leaving little perturbation amplification to correct. At the opposite extreme, when corruption is so strong that the paired views no longer plausibly share one partition, enforcing consistency becomes conceptually inappropriate. SFB-GNN is designed for bounded observation noise between these extremes. This scope is reflected in the stochastic perturbation protocol and in the theoretical assumptions: the framework controls amplification inside a local perturbation neighborhood; it does not claim invariance to arbitrary graph rewrites or to genuine changes in the latent community system.

5.6 Efficiency

Polynomial filtering avoids explicit eigendecomposition. Reconstruction uses negative sampling, drift uses sampled node pairs, and large graphs use neighborhood mini-batches, preventing quadratic memory growth. Training takes $.42$ seconds per epoch on Cora, 2.21 seconds on Coauthor-CS, 8.74 seconds on OGBN-Arxiv, and 15.38 seconds on Reddit, with peak memory from 1.1 to 9.6 GB in the reported setup. Bayesian sampling adds overhead relative to a deterministic encoder, but the method remains feasible on the evaluated large graphs. Appendix H reports batches, fanouts, sample counts, latency, and complete memory measurements.

6 Discussion and Limitations

The results support a narrow claim: explicit stability control improves robustness to the evaluated stochastic structural and feature perturbations while preserving clean clustering quality. The benefit is not uniform across regimes; it is largest when the spectral gap is weak and boundary nodes are uncertain. That regime dependence is desirable because it indicates that SFB-GNN responds to structural fragility rather than imposing an invariant partition regardless of evidence.

Three limitations remain. First, the mean-field posterior is approximate and may underestimate epistemic uncertainty on extremely sparse or irregular graphs. Second, the current implementation assumes a fixed number of communities. Nonparametric or stability-guided model selection is needed when K is unknown. Third, the paired-view objective assumes a shared latent partition. On a genuinely evolving graph, some assignment changes are real; applying the drift loss unchanged could suppress authentic transitions. Dynamic extensions should combine change-point detection, time-conditioned priors, or explicit transition models so that observation noise is separated from community evolution. These directions connect naturally to temporal clustering and dynamic representation learning (Li et al., 2022; Ahsan et al., 2024; Trivedi et al., 2022).

7 Conclusion

SFB-GNN treats partition stability as a first-class objective in attributed-graph community detection. Bounded Fourier propagation limits perturbation amplification, Bayesian uncertainty identifies vulnerable regions, adaptive gating regulates their messages, and permutation-invariant co-assignment consistency directly constrains partition drift. The combined framework lowers structural and feature-induced drift across diverse graph regimes without sacrificing clean clustering or collapsing representations.

AI Use Statement

Generative AI tools were used to assist with language editing, manuscript organization, LaTeX formatting, and improvements to the clarity and conciseness of the presentation. The research questions, mathematical formulations, methodological design, experimental setup, analyses, tables, figures, scientific claims, and selection and verification of citations were determined by the authors. Generative AI was not used as an independent source of scientific evidence, nor was it used to generate or validate experimental results. The authors reviewed and verified all AI-assisted material and take full responsibility for the accuracy, originality, integrity, and final content of the manuscript.

Ethics Statement

This work presents a methodological framework for perturbation-stable community detection in attributed graphs. It does not involve human-subject experiments, clinical interventions, or direct interaction with individuals. The empirical evaluation uses synthetic benchmarks and established real-world graph datasets for aggregate network analysis. We do not attempt to identify individuals, infer sensitive personal attributes, or make decisions about particular users represented in these networks.

Nevertheless, community-detection and uncertainty-estimation methods may be applied to sensitive social, behavioral, or relational data. Any such deployment should therefore account for privacy, in-

formed data governance, biases in graph construction and sampling, and the potential consequences of erroneous or overconfident community assignments. SFB-GNN is intended as a research framework for studying robust and uncertainty-aware community detection under structural and feature perturbations, rather than as a system for automated high-stakes decision making.

Reproducibility Statement

The main paper provides the core formulation of SFB-GNN, including bounded polynomial spectral filtering, Bayesian graph encoding, epistemic uncertainty estimation, uncertainty-aware adaptive message gating, permutation-invariant co-assignment regularization, and the overall learning objective. The Supplementary Material provides the complete graph formulation, perturbation-to-assignment analysis, model and optimization details, training procedure, computational-complexity analysis, and pseudocode.

The Supplementary Material also reports dataset statistics and preprocessing procedures, perturbation-generation protocols, baseline configurations, hyperparameter settings and search ranges, implementation details, statistical-testing procedures, computational-resource measurements, and complete experimental results. The evaluation includes multiple independent runs with different random seeds and repeated perturbation realizations, with aggregation and significance-testing procedures specified in the experimental protocol. Software and hardware settings are documented to support implementation and evaluation. Together, these materials provide the information required to reproduce the model, training procedure, perturbation experiments, and empirical analyses reported in this work.

References

Emmanuel Abbe. Community detection and stochastic block models: recent developments. *Journal of Machine Learning Research*, 18(177):1–86, 2018.

T. Ahsan, J. Kim, and S. Park. Temporal and heterogeneous graph neural networks: A unified review. *Information Fusion*, 115:102115, 2024.

Uri Alon and Eran Yahav. On the bottleneck of graph neural networks and its practical implications. In *International Conference on Learning Representations*, 2021.

Ekin D. Bozdag et al. Adversarial robustness of graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

François-Xavier Briol and Mark Girolami. Bayesian computation with approximate likelihoods. *Statistical Science*, 36(4):528–546, 2021.

Qingfeng Chen, Shiyuan Li, Yixin Liu, Shirui Pan, Geoffrey I Webb, and Shichao Zhang. Uncertainty-aware graph neural networks: A multihop evidence fusion approach. *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

Tianlong Chen, Kaixiong Zhou, Keyu Duan, Wenqing Zheng, Peihao Wang, Xia Hu, and Zhangyang Wang. Bag of tricks for training deeper graph neural networks: A comprehensive benchmark study. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):2769–2781, 2022.

Ke Cheng, Linzhi Peng, Pengyang Wang, Junchen Ye, Leilei Sun, and Bowen Du. Dygkt: Dynamic graph learning for knowledge tracing. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 409–420, 2024.

Mark Cheung, John Shi, Oren Wright, Lavendar Y Jiang, Xujin Liu, and José MF Moura. Graph signal processing and deep learning: Convolution, pooling, and topology. *IEEE Signal Processing Magazine*, 37(6):139–149, 2020.

Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in Neural Information Processing Systems*, 2016.

Ocheme Anthony Ekle and William Eberle. Anomaly detection in dynamic graphs: A comprehensive survey. *ACM Transactions on Knowledge Discovery from Data*, 18(8):1–44, 2024.

S. Ennadir, Y. Abbahaddou, J. F. Lutzeyer, M. Vazirgiannis, and H. Boström. A simple and yet fairly effective defense for graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 21063–21071, 2024.

Farzad V Farahani, Waldemar Karwowski, and Nichole R Lighthall. Application of graph theory for identifying connectivity patterns in human brain networks: a systematic review. *frontiers in Neuroscience*, 13:585, 2019.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Marie-Zahra Alaya, Aurélien Boisbunon, Sébastien Chambon, et al. POT: Python Optimal Transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.

Simon Geisler, Daniel Zügner, and Stephan Günnemann. Reliable graph neural networks via robust aggregation. *Advances in neural information processing systems*, 33:13272–13284, 2020.

Benyamin Ghojogh, Mark Crowley, Fakhri Karray, and Ali Ghodsi. *Elements of dimensionality reduction and manifold learning*. Springer, 2023.

Ksenia Guseva, Sean Darcy, Eva Simon, Lauren V Alteio, Alicia Montesinos-Navarro, and Christina Kaiser. From diversity to complexity: Microbial networks in soils. *Soil Biology and Biochemistry*, 169:108604, 2022.

Maxime Haddouche, Benjamin Guedj, Omar Rivasplata, and John Shawe-Taylor. Pac-bayes unleashed: Generalisation bounds with unbounded losses. *Entropy*, 23(10):1330, 2021.

William L. Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, 2017.

Jesse He, Akbar Rafiey, Gal Mishne, and Yusu Wang. Explaining gnn explanations with edge gradients. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 884–895, 2025.

Jian He, Liang Qiao, Junlin Li, Zhouping Lu, Jiafang Fu, Xiangchun Chen, Xiulin Zhang, and Xulin Hu. Advanced strategies for 3d-printed neural scaffolds: Materials, structure, and nerve remodeling. *Bio-Design and Manufacturing*, 7(5):747–770, 2024.

T. Hoefler, T. Ben-Nun, N. Dryden, et al. Sparsity in deep learning: Pruning and growth for efficient inference and training. *Journal of Machine Learning Research*, 22(241):1–53, 2021.

Abiodun M Ikotun, Absalom E Ezugwu, Laith Abualigah, Belal Abuhaija, and Jia Heming. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, 622:178–210, 2023.

Janggun Jeon, Junho Ahn, and Namgi Kim. Effects of scale regularization in fraud detection graphs. *Electronics*, 14(18):3660, 2025.

Ziheng Jiao, Hongyuan Zhang, and Xuelong Li. Cnn2gnn: How to bridge cnn with gnn. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Ming Jin, Huan Yee Koh, Qingsong Wen, Daniele Zambon, Cesare Alippi, Geoffrey I. Webb, Irwin King, and Shirui Pan. A survey on graph neural networks for time series: Forecasting, classification, imputation, and anomaly detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Bing-Yi Jing, Ting Li, Zhongyuan Lyu, and Dong Xia. Community detection on mixture multilayer networks via regularized tensor decomposition. *The Annals of Statistics*, 49(6):3181–3205, 2021.

Marko Jusup, Petter Holme, Kiyoshi Kanazawa, Misako Takayasu, Ivan Romić, Zhen Wang, Sunčana Geček, Tomislav Lipić, Boris Podobnik, Lin Wang, Wei Luo, Tin Klanjšček, Jingfang Fan, Stefano Boccaletti, and Matjaž Perc. Social physics. *Physics Reports*, 948:1–148, 2022. doi: 10.1016/j.physrep.2021.10.001.

Pavel Kurasov. *Spectral geometry of graphs*. Springer Nature, 2024a.

Pavel Kurasov. *Spectral geometry of graphs*. Springer Nature, 2024b.

Ming Li, Run-Ran Liu, Linyuan Lü, Mao-Bin Hu, Shuqi Xu, and Yi-Cheng Zhang. Percolation on complex networks: Theory and application. *Physics reports*, 907:1–68, 2021.

Rui Li et al. Evolutionary clustering for dynamic networks: A unified framework. *IEEE Transactions on Knowledge and Data Engineering*, 34(5):2361–2375, 2022.

Wenxin Li, Xiao Song, and Yuchun Tu. Graphdrl: Gnn-based deep reinforcement learning for interactive recommendation with sparse data. *Expert Systems with Applications*, 273:126832, 2025.

Yibo Li, Xiao Wang, Yujie Xing, Shaohua Fan, Ruijia Wang, Yaoqi Liu, and Chuan Shi. Graph fairness learning under distribution shifts. In *Proceedings of the ACM Web Conference 2024*, pp. 676–684, 2024.

Han Lin and Wenhe Liu. Symmetry-aware causal-inference-driven web performance modeling: A structure-aware framework for predictive analysis and actionable optimization. *Symmetry*, 17(12): 2058, 2025.

Weisheng Lu and Chimay J. Anumba. *Routledge Handbook of Smart Built Environment*. Routledge, London, UK, 2025.

Costas Mavromatis and George Karypis. Gnn-rag: Graph neural retrieval for efficient large language model reasoning on knowledge graphs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 16682–16699, 2025.

Adrià Mompó Alepuz, Dimitrios Papageorgiou, and Silvia Tolu. Brain-inspired biomimetic robot control: A review. *Frontiers in Neurorobotics*, 18:1395617, 2024.

Weijie Mou, Liwei Wang, and Xinyu Zhai. Generalization bounds of sgd for convex loss functions. *Journal of Machine Learning Research*, 18:1–49, 2017.

Chaoxi Niu, Guansong Pang, and Ling Chen. Graph-level anomaly detection via hierarchical memory networks. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 201–218. Springer, 2023.

Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z Ren, and Anirudha Majumdar. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *ACM Computing Surveys*, 2025.

Clara Stegehuis, Remco Van Der Hofstad, and Johan SH Van Leeuwaarden. Epidemic spreading on complex networks with community structures. *Scientific reports*, 6(1):29748, 2016.

Xing Su, Shan Xue, Fanzhen Liu, Jia Wu, Jian Yang, Chuan Zhou, Wenbin Hu, Cecile Paris, Surya Nepal, Di Jin, et al. A comprehensive survey on community detection with deep learning. *IEEE transactions on neural networks and learning systems*, 35(4):4682–4702, 2022.

Xing Su, Shan Xue, Fanzhen Liu, Jia Wu, Jian Yang, Chuan Zhou, Wenbin Hu, Cecile Paris, Surya Nepal, Di Jin, Quan Z. Sheng, and Philip S. Yu. A comprehensive survey on community detection with deep learning. *IEEE Transactions on Neural Networks and Learning Systems*, 35(4):4682–4702, 2024. doi: 10.1109/TNNLS.2021.3137396.

Rakshit Trivedi, Mehrdad Farajtabar, and Prasenjeet Biswal. Learning representations over dynamic graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 4543–4551, 2022.

Theresa Ullmann, Christian Hennig, and Anne-Laure Boulesteix. Validation of cluster analysis results on validation data: A systematic framework. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(3):e1444, 2022.

Xiao Wang, Meiqi Zhu, Deyu Bo, Peng Cui, Chuan Shi, and Jian Pei. Am-gcn: Adaptive multi-channel graph convolutional networks. In *Proceedings of the 26th ACM SIGKDD International conference on knowledge discovery & data mining*, pp. 1243–1253, 2020.

Yue Yang, Lun Hu, Guodong Li, Dongxu Li, Pengwei Hu, and Xin Luo. Fmvpci: A multiview fusion neural network for identifying protein complex via fuzzy clustering. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 55(9):6189–6202, 2025a. doi: 10.1109/TSMC.2025.3578348.

Yue Yang, Lun Hu, Guodong Li, Dongxu Li, Pengwei Hu, and Xin Luo. Link-based attributed graph clustering via approximate generative bayesian learning. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 55(8):5730–5743, 2025b. doi: 10.1109/TSMC.2025.3572738.

Zhenyu Yang, Ge Zhang, Jia Wu, Jian Yang, Quan Z. Sheng, Shan Xue, Chuan Zhou, Charu C. Aggarwal, H. Peng, Wei Hu, Edwin R. Hancock, and Pietro Lio. State of the art and potentialities of graph-level learning. *ACM Computing Surveys*, 57(2):1–40, 2024.

Kaixiong Zhou, Yuxiao Dong, Kaixin Wang, Wee Sun Lee, Bryan Hooi, Huan Xu, and Jiashi Feng. Understanding and resolving performance degradation in deep graph convolutional networks. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management*, pp. 2728–2737. ACM, 2021.

Raed Abu Zitar, Amani Mohsen, Amal ElFallah Seghrouchni, Frederic Barbaresco, and Nidal A Al-Dmour. Intensive review of drones detection and tracking: Linear kalman filter versus nonlinear regression, an analysis case. *Archives of computational methods in engineering*, 30(5), 2023.