STABILITY-BASED ADVERSARIAL GENERALIZATION BOUNDS FOR HYPERGRAPH CONVOLUTIONAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Hypergraph convolutional networks (HGCNs) propagate information through hyperedges whose learnable weights also determine degree normalization. Analyzing adversarial generalization therefore requires tracking both perturbations of the propagation operator and the coupled updates of layer matrices and hyperedge weights. Under the stated boundedness, Lipschitz activation, and uniform joint-gradient smoothness assumptions, we derive stability-based adversarial generalization bounds for HGCNs trained by two-stream stochastic gradient descent (SGD). We first derive an operator perturbation bound accounting for incidence and weight changes and their effects on normalization. On the explicit family of one large hyperedge and hub-sharing pair edges, over a common weight domain, the ratio between the optimal global hyperedge-weight Lipschitz constants of native propagation and loopless clique expansion grows as the square root of the maximum hyperedge cardinality. Propagating operator sensitivity through the coupled SGD dynamics yields structural and feature adversarial generalization bounds that expose the effects of degree imbalance, network depth, and accumulated optimization expansion. For T SGD updates and n training examples, the resulting bounds scale as $\mathcal{O}(T/n)$ for a fixed architecture when the stated regularity and step-size conditions hold and the loss, gradient, and accumulated-expansion terms are bounded independently of n and T. A separate one-layer construction under zero attacks, zero regularization, and a restricted training horizon gives a uniform-stability lower bound of order $r_{deg}\eta T/n$, where r_{deg} measures degree imbalance and η is the common step size. Sample-size experiments show both lower gaps at larger training sizes and feature-attack reversals on Cora and PubMed. Controlled forward passes attain the predicted geometric depth dependence.

1 INTRODUCTION

Higher-order interactions motivate hypergraph convolutional networks (HGCNs), which propagate information through hyperedges (Zhou et al., 2006; Feng et al., 2019). Learning the hyperedge weight matrix W_e changes both the incidence product and weighted node degrees, affecting every incident node through normalization and subsequent layers. Sensitivity depends on incident degrees and perturbation direction: uniform positive rescaling leaves the native operator unchanged, whereas relative weight changes can alter propagation. The challenge is to connect this sensitivity to stability and adversarial generalization under joint training of layer matrices and hyperedge weights.

To address this challenge, we build on graph convolutional network (GCN) stability analysis (Verma & Zhang, 2019; Zhou & Wang, 2021; Cao et al., 2025). A telescoping decomposition of the normalized operator bounds joint incidence and weight perturbations. We propagate this sensitivity across layers and track the coupled parameter updates through two-stream stochastic gradient descent (SGD). Directional sensitivity measurements on an explicit hypergraph family support the predicted separation between native propagation and loopless clique expansion. Sample-size experiments show decreasing gaps in many conditions and feature-attack reversals on Cora and PubMed.

Contributions. Our main contributions can be described as follows.

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.

- **Operator separation on a constructed family**. On an explicit family with $m = d_{e,\max}$(the maximum hyperedge cardinality), matching bounds establish optimal global hyperedge-weight Lipschitz constants $\Theta(m)$ for native propagation and $\Theta(\sqrt{m})$ for loopless clique expansion. Their ratio is $\Theta(\sqrt{m})$ over a common weight domain and diagonal spectral-norm budget.
- **Two-stream stability and adversarial generalization.** Under Assumptions A1–A4 and the step-size condition, Theorems 2 and 3 bound the expected structural and feature adversarial gaps through degree imbalance, depth, and accumulated expansion. For a fixed architecture, the bounds are $\mathcal{O}(T/n)$ when loss and gradient envelopes, step sizes, and accumulated expansion are bounded independently of sample size n and update count T.
- **Local instability and Cheeger scaling.** Proposition 1 establishes a one-layer uniform-stability lower bound of order $r_{deg}\eta T/n$ under its local horizon, zero attacks, and no regularization. Here r_{deg} is the degree-imbalance ratio in Equation (2) and η the common step size. Degree and weight floors vary across the family. A separate hub-leaf family has Cheeger constant $\Theta(1/r_{deg})$ (Proposition 4).

2 Related Work

Propagation and stability. Spectral hypergraph models use incidence-based propagation (Zhou et al., 2006; Feng et al., 2019). Algorithmic stability links the effect of replacing a training example to an expected risk gap (Hardt et al., 2016). Clean graph neural network (GNN) analysis track propagation-dependent constants (Verma & Zhang, 2019; Zhou & Wang, 2021), while adversarial-training analysis account for inner maximization in the update dynamics (Farnia & Özdağlar, 2021; Xing et al., 2021; Xiao et al., 2022). Multi-layer GCN analysis considers graph propagation under adversarial training (Cao et al., 2025). When hyperedge weights and layer matrices are learned jointly, hyperedge-weight updates also change the normalized propagation operator. The resulting coupling enters our stability analysis (Table 6, Appendix G).

Robustness certification. Randomized smoothing certifies predictions of fixed graph and hypergraph models (Bojchevski et al., 2020; Geisler et al., 2020; Zhang et al., 2024). Collective certification uses node dependencies to certify sets of predictions (Schuchardt et al., 2023). We bound the absolute expected population–empirical adversarial risk difference after training. Appendix G also covers capacity-based bounds and empirical defenses.

3 Preliminaries and Stability Framework

3.1 Hypergraph and HGCN Architecture

Hypergraph and notation. A weighted hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E}, W_e)$ has $N = |\mathcal{V}|$ nodes and $M = |\mathcal{E}|$ hyperedges. Writing $\mathcal{E} = \{e_1, \ldots, e_M\}$, its learnable weights are $W_e = \mathrm{diag}(w_{e_1}, \ldots, w_{e_M}) \in \mathbb{R}^{M\times M}$. Its incidence matrix $H \in \{0,1\}^{N\times M}$ satisfies $H_{ve} = 1$ if and only if $v \in e$. Hyperedge cardinalities $d_e = \sum_v H_{ve}$ and weighted node degrees $d_v = \sum_e H_{ve} w_e$ give $D_e = \mathrm{diag}(d_e)$ and $D_v = \mathrm{diag}(d_v)$. Their common lower and upper bounds are $d_{e,\min}, d_{e,\max}$ and $d_{v,\min}, d_{v,\max}$ under Assumption A1. Features are $X \in \mathbb{R}^{N\times d_{in}}$, with input dimension d_{in} and n labeled training examples. Matrix $\|\cdot\|_2$ denotes the spectral norm, vector $\|\cdot\|_2$ the Euclidean norm, and $\|\cdot\|_F$ the Frobenius norm.

Layer-wise propagation. For a K-layer HGCN, let $d_0 = d_{in}$ and let d_ℓ be the representation dimension at layer $\ell = 1, \ldots, K$. The learnable layer matrices are $\mathcal{W} = \{W^{(0)}, \ldots, W^{(K-1)}\}$, where $W^{(\ell-1)} \in \mathbb{R}^{d_{\ell-1}\times d_\ell}$. At each layer, $\sigma : \mathbb{R} \to \mathbb{R}$ denotes the entrywise activation used at that layer. Hidden layers may use ReLU, and the final layer uses the identity. Assumption A3 requires a common bound v_σ across all layers. Let $\mathbf{P} \in \mathbb{R}^{N\times N}$ denote the **hypergraph propagation operator** defined as (Zhou et al. (2006)):

$$\mathbf{P} = D_v^{-1/2} H W_e D_e^{-1} H^\top D_v^{-1/2}. \tag{1}$$

Starting from $X^{(0)} = X$, the representations $X^{(\ell)} \in \mathbb{R}^{N\times d_\ell}$ satisfy

$$X^{(\ell)} = \sigma(\mathbf{P} X^{(\ell-1)} W^{(\ell-1)}). \qquad \ell = 1, \ldots, K,$$

The readout g returns one prediction per node, giving $f(\mathbf{P}, \mathcal{W}, X) = g(X^{(K)})$. To express the degree dependence of the operator and gradient bounds, define

$$r_{deg} = \frac{\sqrt{d_{v,\max} d_{e,\max}}}{d_{v,\min} d_{e,\min}}. \tag{2}$$

Its value depends on degree imbalance and the scale of hyperedge weights. We use a hat to denote the compared counterpart of a quantity. It represents the perturbed quantity in attack comparisons and the neighboring trajectory in stability comparisons. For incidence matrices, $\|\cdot\|_\infty$ and $\|\cdot\|_1$ denote the maximum absolute row and column sums, respectively. For diagonal W_e, the spectral and induced ∞ norms coincide. Appendix A relates the entrywise and spectral feature budgets.

3.2 Threat Models and Generalization Objective

With the propagation model and notation fixed, we now specify the structural and feature threat models and the corresponding adversarial risks. Let D be the supervision distribution conditional on public H, X, and let $z = (x, y)$ denote a supervision example. For an example at node v, $x = X_{v,:} \in \mathbb{R}^{d_{in}}$ and y is its label. The example retains the node identity, which is suppressed in the notation $z = (x, y)$. The inputs are the incidence matrix H, feature matrix X, and training sample $S = \{z_i = (x_i, y_i)\}_{i=1}^n \sim D^n$. Let $\mathcal{L}(\hat{y}, y)$ denote the clean loss for prediction $\hat{y}$ and label y. For the example at node v, we write $f(\mathbf{P}, \mathcal{W}, x) = [g(X^{(K)})]_{v,:}$. For $\Delta x \in \mathbb{R}^{d_{in}}$, the vector norm is $\|\Delta x\|_\infty = \max_{1 \le j \le d_{in}} |\Delta x_j|$. Feature attacks replace the feature at node v by $x + \Delta x$. For $\Delta H \in \mathbb{R}^{N \times M}$ and prescribed budgets $\epsilon_H, \epsilon_x \ge 0$, the structural and feature adversarial losses at fixed parameters are

$$\tilde{\mathcal{L}}_{\mathrm{str}}(f(\mathbf{P}, \mathcal{W}, x), y) = \sup_{\max\{\|\Delta H\|_\infty, \|\Delta H\|_1\} \le \epsilon_H} \mathcal{L}(f(\mathbf{P}(H + \Delta H, W_e), \mathcal{W}, x), y), \tag{3}$$

$$\tilde{\mathcal{L}}_{\mathrm{fea}}(f(\mathbf{P}, \mathcal{W}, x), y) = \sup_{\|\Delta x\|_\infty \le \epsilon_x} \mathcal{L}(f(\mathbf{P}(H, W_e), \mathcal{W}, x + \Delta x), y). \tag{4}$$

Structural perturbations additionally satisfy $H + \Delta H \ge 0$ entrywise, with positive row and column sums so that $\mathbf{P}$ is defined. This continuous incidence model uses column sums as d_e. Choose the loss $\tilde{\mathcal{L}} \in \{\tilde{\mathcal{L}}_{\mathrm{str}}, \tilde{\mathcal{L}}_{\mathrm{fea}}\}$ for the selected threat model. Given regularization coefficient $\lambda \ge 0$, the optimization target is

$$\inf_{\mathcal{W}, W_e} \left\{ \frac{1}{n} \sum_{i=1}^n \tilde{\mathcal{L}}(f(\mathbf{P}, \mathcal{W}, x_i), y_i) + \lambda \|W_e\|_F^2 \right\}, \tag{5}$$

over the layer matrices $\mathcal{W}$ and diagonal W_e with $w_e > 0$. Hyperedge weights are outer optimization variables, while each inner supremum changes only incidence or features. For comparisons of weight matrices, we retain the budget $\epsilon_W = \|W_e - \widehat{W}_e\|_\infty$ (equal to $\|W_e - \widehat{W}_e\|_2$ for diagonal W_e). Section 3.4 specifies maximizer and gradient regularity.

Let $\mathcal{A}_S = (\mathcal{W}_S, W_{e,S})$ be the output of algorithm $\mathcal{A}$, and let $S^{(i)}$ replace example i by an independent draw from D. Both runs use the same public inputs H, X, fixed independently of the training sample. For $z = (x, y)$, we write the adversarial loss of the trained model as $\tilde{\mathcal{L}}(\mathcal{A}_S, z) = \tilde{\mathcal{L}}(f(\mathbf{P}(H, W_{e,S}), \mathcal{W}_S, x), y)$. The evaluated adversarial losses are assumed integrable over the data and algorithmic randomness, as required by Proposition 2. The population and empirical adversarial risks are

$$\mathcal{R}(\mathcal{A}_S) = \mathbb{E}_{z \sim D}[\tilde{\mathcal{L}}(\mathcal{A}_S, z)], \qquad \mathcal{R}_{\mathrm{emp}}(\mathcal{A}_S) = \frac{1}{n} \sum_{i=1}^n \tilde{\mathcal{L}}(\mathcal{A}_S, z_i).$$

Both risks exclude the training regularizer. The expected adversarial generalization gap is

$$\mathrm{Gap} = \left| \mathbb{E}_{S, \mathcal{A}}[\mathcal{R}(\mathcal{A}_S) - \mathcal{R}_{\mathrm{emp}}(\mathcal{A}_S)] \right|. \tag{6}$$

3.3 Uniform Stability and Two-Stream SGD

We control this expected gap through uniform stability of the two-stream SGD algorithm. For $\beta \ge 0$, an algorithm is β-uniformly stable for the selected adversarial loss if, for every pair $S, S^{(i)}$ differing

in one example and with z ranging over the supervision domain of D,

$$\sup_z \left| \mathbb{E}_{\mathcal{A}}[\tilde{\mathcal{L}}(\mathcal{A}_S, z) - \tilde{\mathcal{L}}(\mathcal{A}_{S^{(i)}}, z)] \right| \leq \beta.$$

Expectation is over algorithmic randomness (Hardt et al., 2016). Proposition 2 establishes Gap $\leq \beta$.

Definition 1 (Two-Stream SGD)**.** *Given H, X, S, a selected adversarial loss, and λ, initialize $(\mathcal{W}_1, W_{e,1})$ at $t = 1$ within the parameter domain of Section 3.2. For a prescribed number of updates T and step sizes $\eta_t, \xi_t > 0$, sample i_t independently and uniformly from $\{1, \ldots, n\}$ at each $t = 1, \ldots, T$. With $\mathbf{P} = \mathbf{P}(H, W_{e,t})$, two-stream SGD applies the following updates for $\ell = 0, \ldots, K-1$ to the objective in Equation* (5)*:*

$$W_{t+1}^{(\ell)} = W_t^{(\ell)} - \eta_t \nabla_{W^{(\ell)}} \tilde{\mathcal{L}}(f(\mathbf{P}, \mathcal{W}_t, x_{i_t}), y_{i_t}), \tag{7}$$

$$W_{e,t+1} = (1 - 2\xi_t \lambda) W_{e,t} - \xi_t \nabla_{W_e} \tilde{\mathcal{L}}(f(\mathbf{P}, \mathcal{W}_t, x_{i_t}), y_{i_t}), \tag{8}$$

Both gradients are evaluated at the same pre-update pair $(\mathcal{W}_t, W_{e,t})$. Gradients with respect to W_e act on its diagonal coordinates, preserving the diagonal parameterization. The output after T updates is $\mathcal{A}_S = (\mathcal{W}_{T+1}, W_{e,T+1})$. For the stability comparison, neighboring runs share initialization and sampled indices. Their trajectory conditions are stated separately in Section 3.4.

3.4 Regularity Assumptions A1–A4

All bounds below hold uniformly over neighboring trajectories, admissible attacks, and intermediate parameter pairs and paths used in the proofs. Appendix C.1 discusses their scope.

- **A1 (Bounded degrees and weights).** For every node $v \in \mathcal{V}$ and hyperedge $e \in \mathcal{E}$, $0 < d_{v,\min} \leq d_v \leq d_{v,\max}$, $1 \leq d_{e,\min} \leq d_e \leq d_{e,\max}$, and $0 < w_{\min} \leq w_e \leq w_{\max} < \infty$. Structural attacks use nonnegative incidence matrices and a common, parameter-independent feasible set satisfying these bounds for every compared parameter pair.
- **A2 (Bounded parameters and inputs).** With norm bounds $C_W, C_{W_e}, C_x, C_{\mathbf{P}}$, $\|W^{(\ell)}\|_2 \leq C_W$, $\|W_e\|_2 \leq C_{W_e}$, $\|X\|_2 \leq C_x$, $\|\mathbf{P}\|_2 \leq C_{\mathbf{P}}$, for $\ell = 0, \ldots, K-1$. The readout g has Lipschitz constant C_g.
- **A3 (Spectral-norm Lipschitz activation).** For each $\ell = 1, \ldots, K$ and all admissible preactivations $U, V \in \mathbb{R}^{N \times d_\ell}$, a common constant v_σ bounds the activation used at each layer:
$$\|\sigma(U) - \sigma(V)\|_2 \leq v_\sigma \|U - V\|_2, \qquad \sigma(0) = 0.$$
- **A4 (Lipschitz and smooth loss).** At a fixed admissible perturbation, write $\mathcal{L}(\mathcal{W}, W_e, z) = \mathcal{L}(f(\mathbf{P}, \mathcal{W}, x), y)$ for $z = (x, y)$, and use $\tilde{\mathcal{L}}(\mathcal{W}, W_e, z)$ for the selected adversarial loss. For layer collections, $\|\cdot\|_2$ denotes the sum of component spectral norms: $\|\mathcal{W} - \widehat{\mathcal{W}}\|_2 = \sum_{\ell=0}^{K-1} \|W^{(\ell)} - \widehat{W}^{(\ell)}\|_2$, $\|\nabla_{\mathcal{W}} \tilde{\mathcal{L}}\|_2 = \sum_{\ell=0}^{K-1} \|\nabla_{W^{(\ell)}} \tilde{\mathcal{L}}\|_2$. At SGD evaluation points, the adversarial objective is differentiable and its gradient is represented by a selected maximizing loss. Let v_W denote the loss Lipschitz constant, including the readout, and let μ_W, μ_{W_e} denote the threat-specific joint gradient Lipschitz constants. For any two compared parameter pairs at the same z, we require the following inequalities, with the perturbation held fixed in Equation (9):
$$\left|\mathcal{L}(\mathcal{W}, W_e, z) - \mathcal{L}(\widehat{\mathcal{W}}, \widehat{W}_e, z)\right| \leq v_W \|X^{(K)}(\mathcal{W}, W_e) - X^{(K)}(\widehat{\mathcal{W}}, \widehat{W}_e)\|_2, \tag{9}$$
$$\left\|\nabla_{\mathcal{W}} \tilde{\mathcal{L}}(\mathcal{W}, W_e, z) - \nabla_{\mathcal{W}} \tilde{\mathcal{L}}(\widehat{\mathcal{W}}, \widehat{W}_e, z)\right\|_2 \leq \mu_W \left(\|\mathcal{W} - \widehat{\mathcal{W}}\|_2 + \|W_e - \widehat{W}_e\|_\infty\right), \tag{10}$$
$$\left\|\nabla_{W_e} \tilde{\mathcal{L}}(\mathcal{W}, W_e, z) - \nabla_{W_e} \tilde{\mathcal{L}}(\widehat{\mathcal{W}}, \widehat{W}_e, z)\right\|_2 \leq \mu_{W_e} \left(\|W_e - \widehat{W}_e\|_\infty + \|\mathcal{W} - \widehat{\mathcal{W}}\|_2\right). \tag{11}$$

Under Assumption A1, Equation (1) has $C_{\mathbf{P}} = 1$. Assumption A2 imposes uniform bounds along all compared training trajectories, in addition to the prescribed weight decay. Together with Assumption A3, these bounds give $\|X^{(\ell)}\|_2 \leq C_x (v_\sigma C_{\mathbf{P}} C_W)^\ell$. For entrywise ReLU, Appendix C.1 gives a sufficient spectral-norm Lipschitz constant. Assumption A4 controls changes in the parameter gradients, including those caused by changes in the selected adversarial maximizer.

3.5 KEY COMPOSITE QUANTITIES

Define the one-step expansion factor

$$\rho_t = \max\{1 + \eta_t \mu_W,\ 1 - \xi_t(2\lambda - \mu_{W_e})\} + \xi_t \mu_{W_e}. \tag{12}$$

For $0 \le t \le T$, define its product over updates $t+1$ through T by

$$\Psi_{t,T} = \prod_{s=t+1}^{T} \rho_s, \qquad \Psi_{T,T} = 1. \tag{13}$$

For a constant hyperedge step size $\xi_t = \xi$, also define the isolated hyperedge factor $\rho_{W_e} = 1 - \xi(2\lambda - \mu_{W_e})$ and the factor $\rho_{\text{imp}} = v_\sigma C_{\mathbf{P}} C_W$ governing the representation bound across layers. Lemma 2 and Appendix B establish the roles of these quantities in the bounds.

4 MAIN RESULTS

We first compare native propagation and loopless clique expansion without training dynamics. We then use the expansion factors above to derive stability-based generalization bounds for joint training of layer matrices and hyperedge weights.

4.1 GLOBAL OPERATOR SEPARATION ON A CONSTRUCTED FAMILY

We compare native propagation and loopless clique expansion on the following hypergraph family, using a common weight domain and perturbation norm. The proof is given in Appendix D.1.

Theorem 1 (Global Separation on an Explicit Family)**.** *For integers $m \ge 4$, let $\mathcal{H}_m$ have vertices $0, \dots, m-1$, one hyperedge $e_0 = \{0, \dots, m-1\}$, and pair hyperedges $e_j = \{0, j\}$ for $1 \le j < m$. Their weights vary independently over*

$$\mathcal{B}_m = \left\{ \operatorname{diag}(w_0, \dots, w_{m-1}) : \tfrac{1}{2} \le w_0 \le 2, \quad \tfrac{1}{2m} \le w_j \le \tfrac{2}{m}\ (j \ge 1) \right\}.$$

Let $\mathcal{G}^{CE}$ replace each hyperedge e by a loopless clique with pair weight w_e/d_e. Write $\mathbf{P}^{CE} = (D_v^{CE})^{-1/2} A^{CE} (D_v^{CE})^{-1/2}$, where A^{CE} and D_v^{CE} are its adjacency and degree matrices, and set $\mathbf{P}^{HG} = \mathbf{P}$. Define the optimal global constants by

$$C_2^Q(\mathcal{H}_m) = \sup_{\substack{W_e, \widehat{W}_e \in \mathcal{B}_m \\ W_e \ne \widehat{W}_e}} \frac{\|\mathbf{P}^Q(W_e) - \mathbf{P}^Q(\widehat{W}_e)\|_2}{\|W_e - \widehat{W}_e\|_2}, \qquad Q \in \{HG, CE\}. \tag{14}$$

Then, with constants independent of m,

$$C_2^{HG}(\mathcal{H}_m) = \Theta(m), \quad C_2^{CE}(\mathcal{H}_m) = \Theta(\sqrt{m}), \quad \kappa(\mathcal{H}_m) := \frac{C_2^{HG}(\mathcal{H}_m)}{C_2^{CE}(\mathcal{H}_m)} = \Theta(\sqrt{d_{e,\max}}). \tag{15}$$

Remark 1 (Mechanism and scope)**.** *The separation is driven by different directional scalings at the hub. Coordinated changes in the pair-edge weights produce a directional derivative of order m in the native operator's hub diagonal entry, whereas the corresponding hub row of the loopless clique operator has derivative norm of order $\sqrt{m}$. Together with uniform upper bounds on $\mathcal{B}_m$, these directional lower bounds yield the global separation. Corollary 1 gives the corresponding exact directional asymptotics.*

The result concerns the optimal global operator constants on $\mathcal{B}_m$, where $w_{\min} = 1/(2m)$ while node degrees remain uniformly positive. These constants are distinct from the general perturbation upper bound C_2 in Lemma 1. Their implications for gradient bounds and generalization also depend on the remaining prefactors and the training dynamics (see Remark 6).

4.2 STRUCTURAL GENERALIZATION BOUND

Theorem 1 compares operators on a constructed family without training. For structural attacks during joint training, incidence and weight changes also affect degree normalization. Lemma 1 bounds the resulting operator change, and Lemma 2 uses gradient bounds to track parameter differences between neighboring SGD runs. Their proofs appear in Appendices C.3 and C.6.

Lemma 1 (Operator perturbation). *Assume that both (H, W_e) and $(\widehat{H}, \widehat{W}_e)$ satisfy Assumption A1 with the same degree and weight bounds. Let $\Delta H = H - \widehat{H}$ and $\Delta W_e = W_e - \widehat{W}_e$. Suppose $\|\Delta H\|_\infty \le \epsilon_H$, $\|\Delta H\|_1 \le \epsilon_H$, and $\|\Delta W_e\|_2 \le \epsilon_W$. Then*

$$\|\mathbf{P} - \widehat{\mathbf{P}}\|_2 \le C_1 \epsilon_H + C_2 \epsilon_W,$$

where

$$C_1 = w_{\max} r_{deg} \left(\frac{2}{\sqrt{w_{\min}}} + \frac{w_{\max} d_{e,\min} + d_{v,\min}}{w_{\min}} \, r_{deg} \right), \tag{16}$$

$$C_2 = \frac{d_{e,\min} \, r_{deg}^2}{w_{\min}} \left(d_{v,\min} + \frac{w_{\max} d_{v,\max}}{w_{\min}} \right). \tag{17}$$

The row- and column-sum budgets control changes in weighted node degrees and hyperedge degrees, respectively. If $\|\Delta H\|_F \le \bar{\epsilon}$, then $\epsilon_H = \sqrt{\max\{N, M\}}\,\bar{\epsilon}$ is valid. Measured row and column sums may give a smaller value (Remark 3). The degree and weight floors control the normalization. Lemmas 5– 6 provide the gradient bounds used next to compare neighboring SGD trajectories.

Lemma 2 (Two-stream SGD stability recursion). *For two-stream SGD trajectories trained on neighboring datasets S and $S^{(i)}$, where $S^{(i)}$ replaces example i in S, define $\Delta_t = \sum_{\ell=0}^{K-1} \|W_t^{(\ell)} - \widehat{W}_t^{(\ell)}\|_2 + \|W_{e,t} - \widehat{W}_{e,t}\|_2$. Under Assumptions A1 through A4, suppose $0 \le 2\xi_t\lambda \le 1$. Let ρ_t be defined as in Equation (12). Let G_W bound the sum of layer-gradient spectral norms and G_{W_e} the hyperedge-gradient spectral norm, uniformly over neighboring trajectories, sampled indices, and admissible maximizers for the selected threat. One explicit choice is:*

$$G_W = K C_x v_W (v_\sigma C_{\mathbf{P}})^K C_W^{K-1}, \tag{18}$$

$$G_{W_e} = K C_x v_W C_2 v_\sigma^K C_{\mathbf{P}}^{K-1} C_W^K. \tag{19}$$

The constants are uniform on the selected threat domain. Their values may differ between structural and feature attacks. Each layer-gradient norm is bounded by G_W/K. Then the expected parameter divergence satisfies

$$\mathbb{E}[\Delta_{t+1}] \le \rho_t \mathbb{E}[\Delta_t] + \frac{2}{n}\big(\eta_t G_W + \xi_t G_{W_e}\big). \tag{20}$$

To bound the generalization gap, Lemma 9 converts parameter divergence into loss stability. Proposition 2 then bounds the expected population–empirical risk difference. Appendix D.3 contains the complete derivation.

Theorem 2 (Adversarial Structural Generalization Gap). *Consider T two-stream SGD updates (Equations (7) and (8)) of a K-layer HGCN on the objective in Equation (5), with $\tilde{\mathcal{L}} = \tilde{\mathcal{L}}_{\mathrm{str}}$. Under Assumptions A1 through A4, suppose $0 \le 2\xi_t\lambda \le 1$ for every t. With G_W/K and G_{W_e} as in Equations (18)–(19), define the parameter-to-loss conversion constants*

$$C_4 = v_W C_x (v_\sigma C_{\mathbf{P}})^K C_W^{K-1}, \qquad C_5 = v_W C_3^{\mathrm{rep}} C_2, \tag{21}$$

where C_3^{rep} is the representation Lipschitz constant with respect to the propagation operator from Lemma 4 (explicit coefficient in Appendix C.4), satisfying $\|X^{(K)} - \widehat{X}^{(K)}\|_2 \le C_3^{\mathrm{rep}} \|\mathbf{P} - \widehat{\mathbf{P}}\|_2$ for fixed $\mathcal{W}$ and X, with $\|\hat{\mathbf{P}}\|_2 \le C_{\mathbf{P}}$. Define the Lipschitz envelope

$$C_{\mathrm{Lip}}^{\mathrm{str}} = \max\big(G_W/K,\, G_{W_e},\, C_4,\, C_5\big).$$

Let $\Psi_{t,T}$ be as in Equation (13). Then the adversarial generalization gap defined in Equation (6) satisfies

$$\mathrm{Gap} \le \frac{2C_{\mathrm{Lip}}^{\mathrm{str}}}{n} \sum_{t=1}^{T} \Big(\eta_t G_W + \xi_t G_{W_e}\Big) \Psi_{t,T}. \tag{22}$$

Its compact form is

$$\mathrm{Gap} \le \frac{2\big(C_{\mathrm{Lip}}^{\mathrm{str}}\big)^2}{n} \sum_{t=1}^{T} (K\eta_t + \xi_t) \Psi_{t,T} \tag{23}$$

Remark 2 (Non-vacuousness and bounded accumulation). *The $1/n$ factor is the probability of sampling the replaced example. For fixed K, Equation (23) is $\mathcal{O}(T/n)$ when $C_{\mathrm{Lip}}^{\mathrm{str}}$, the step sizes, and $\sup_{t,T} \Psi_{t,T}$ are bounded independently of n, T. Assumption A4 controls gradient changes across activation and maximizing-attack states. Depth enters through ρ_{imp}, while $\Psi_{t,T}$ captures optimization accumulation. Appendix B gives the corresponding sufficient controls.*

4.3 Feature Generalization Bound

For feature attacks, H is fixed but learned hyperedge weights still induce trajectory-dependent operators $\mathbf{P}(H, W_e)$, so r_{deg} enters through G_{W_e}. The constants remain uniform on the feature-threat domain, including the A2 bound C_x on attacked feature matrices. Appendix D.4 gives the proof and threat-model comparison, while Appendix C.9 gives the fixed-parameter ReLU Frobenius-norm refinement.

Theorem 3 (Adversarial Feature Generalization Gap)**.** *Consider a K-layer HGCN trained for T two-stream SGD updates (Equations* (7) *and* (8)*) on Equation* (5)*, with $\tilde{\mathcal{L}} = \tilde{\mathcal{L}}_{\mathrm{fea}}$. Assume A1–A4, $0 \le 2\xi_t\lambda \le 1$ for every t, and fixed incidence matrix H. With C_4, C_5 from Equation* (21) *and G_W, G_{W_e} from Equations* (18)– (19)*, define*

$$C_{\mathrm{Lip}}^{\mathrm{fea}} = \max\big(G_W,\, G_{W_e},\, C_4,\, C_5\big).$$

Let $\Psi_{t,T}$ be as in Equation (13)*. The resulting adversarial generalization gap, defined in Equation* (6)*, is bounded by*

$$\mathrm{Gap} \le \frac{2C_{\mathrm{Lip}}^{\mathrm{fea}}}{n} \sum_{t=1}^{T} \big(\eta_t G_W + \xi_t G_{W_e}\big)\Psi_{t,T}. \tag{24}$$

The following compact bound also holds,

$$\mathrm{Gap} \le \frac{2\big(C_{\mathrm{Lip}}^{\mathrm{fea}}\big)^2}{n} \sum_{t=1}^{T} (\eta_t + \xi_t)\Psi_{t,T} \tag{25}$$

Complementary local instability. The next proposition examines degree-dependent replacement sensitivity in a restricted one-layer construction.

Proposition 1 (Local Horizon Hard Instance for r_{deg}-Dependent Stability)**.** *For integers $r \ge 2$ and $n \ge 1$, there exist a weighted hypergraph with initial $r_{deg} = r$, a one-layer HGCN, and neighboring samples $S, S^{(i)}$ of size n satisfying Assumptions A1–A4 on a local parameter domain. With zero attack budgets, $\lambda = 0$, common initialization, and $\eta_t = \xi_t = \eta$, two-stream SGD satisfies*

$$\sup_z \Big|\mathbb{E}_{\mathcal{A}}[\tilde{\mathcal{L}}(\mathcal{A}_S, z) - \tilde{\mathcal{L}}(\mathcal{A}_{S^{(i)}}, z)]\Big| \ge c\,\frac{r_{deg}\eta T}{n}$$

whenever $\eta T r^2 \le c_{\mathrm{loc}}$, for universal $c, c_{\mathrm{loc}} > 0$. The A1 lower bounds may depend on r.

This construction establishes degree-dependent replacement sensitivity for the stated one-layer, zero-attack, unregularized local regime. Appendix E.1 gives the complete construction and a stronger bound for this instance. A separate hub–leaf family relates the same degree statistic to connectivity. Under the weighted cut-ratio convention in Appendix E.2, its Cheeger constant satisfies $h(\mathcal{H}) = \Theta(1/r_{deg})$ (Proposition 4).

5 Experiments

We examine directional operator separation (**RQ1**), sample-size dependence (**RQ2**), degree imbalance and attack response (**RQ3**), and controlled depth dependence (**RQ4**). Appendix F gives the protocols and extended results. In RQ2–RQ3, structural attacks perturb continuous incidence values with Frobenius budget $\bar{\epsilon}$. The GCN baseline applies the corresponding constraint to its adjacency matrix. Appendix F.5 specifies the attack protocols and explains their relation to the theoretical budget ϵ_H.

5.1 RQ1: Directional Operator Separation

We first examine the construction underlying Theorem 1. Varying the large hyperedge's weight tests the $\sqrt{m}$ directional sensitivity ratio in Corollary 1. Table 1 reports the results, and Appendix F.6 describes the protocol.

In Table 1, s_{HG} remains nearly constant while s_{CE} decreases. The ratio rises from 1.746 to 13.120, and $\kappa_{\mathrm{dir}}/\sqrt{m}$ increases from 0.437 to 0.580, approaching the limit $\sqrt{6}/4 \approx 0.612$ in Corollary 1. This deterministic experiment isolates the directional operator mechanism on the constructed family.

Table 1: Directional sensitivities s_{HG} and s_{CE} on the constructed size-m family, estimated by central differences in the large-hyperedge weight with $\epsilon = 10^{-5}$. The ratio $\kappa_{\mathrm{dir}} = s_{HG}/s_{CE}$ uses unrounded values (Appendix F.6).

m	$s_{HG}(m)$	$s_{CE}(m)$	$\kappa_{\mathrm{dir}}(m)$	$\kappa_{\mathrm{dir}}(m)/\sqrt{m}$
16	0.111	0.064	1.746	0.437
32	0.118	0.043	2.755	0.487
64	0.122	0.029	4.182	0.523
128	0.123	0.020	6.203	0.548
256	0.124	0.014	9.066	0.567
512	0.125	0.009	13.120	0.580

5.2 RQ2: Sample-Size Dependence

We examine sample-size dependence at fixed T on Cora and PubMed, using nested training sets with fixed validation and test nodes for each seed. We report $M1_{\mathrm{CE}} = |\widehat{R}^{\mathrm{adv}}_{\mathrm{train}} - \widehat{R}^{\mathrm{adv}}_{\mathrm{test}}|$ for the splitwise cross-entropy gap and $M4 = p_{\mathrm{hit}}(T)\Delta_T^{\mathrm{hit}}$ for the hit-probability-weighted terminal parameter divergence (Appendix F.4).

Figure 1 shows that, under the Cora structural attack, both $M1_{\mathrm{CE}}$ and $M4$ are lower at the largest training size for HGCN and GCN at $K = 2$ and $K = 4$. The extended results in Appendix F.7 show the same endpoint pattern in most clean and structural settings.

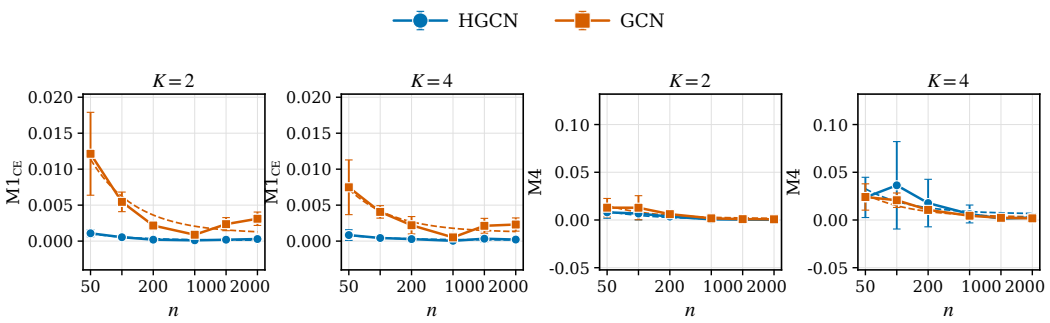

Figure 1: Sample-size dependence on Cora for HGCN and GCN under support-restricted structural PGD with Frobenius budget $\bar{\epsilon} = 10$. Training uses $T = 120$ sampled-SGD updates and ten seeds. The four panels show $M1_{\mathrm{CE}}$ and $M4$ at $K = 2$ and $K = 4$ against log-scaled training size n. Markers and error bars show mean $\pm$ sample standard deviation, and dashed curves show $a/n + b$ fits. Appendix F.5 gives the attack protocol.

Feature attacks produce several exceptions. For Cora HGCN at $K = 2$ and $\epsilon_x = 0.010$, the final $M1_{\mathrm{CE}}$ is higher at the largest training size. PubMed GCN shows a similar reversal at $(K, \epsilon_x) = (2, 0.005)$ and $(4, 0.010)$. At $(K, \epsilon_x) = (4, 0.005)$, PubMed GCN also ends with a higher $M1_{\mathrm{CE}}$. In contrast, all retained $M4$ curves end lower at the largest training size. These reversals may partly reflect the broader all-row feature attack used in the experiments relative to the single-row threat analyzed theoretically.

These experiments characterize finite-split sample-size behavior under the implemented training and attack protocols. The theoretical results concern expected adversarial gaps under A1–A4. The reported feature experiments use all-row PGD. Appendix F.7 gives the detailed correspondence between the experimental quantities and the theoretical setting.

5.3 RQ3: Degree Imbalance and Attack Response

The operator and stability bounds depend explicitly on r_{deg}, motivating an empirical comparison with structural attack response. We use the initial r_{deg} at unit hyperedge weights and $M2 = \widehat{R}_{\mathrm{test}}(\widehat{H}) -$

$\widehat{R}_{\text{test}}(H)$, the attacked-minus-clean test cross-entropy at fixed trained parameters, with $\|\widehat{H} - H\|_F = 5$ and zero weight-attack budget.

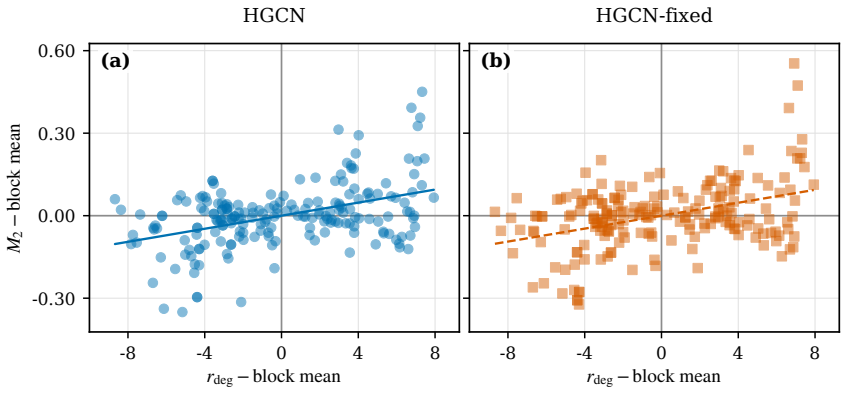

Figure 2: Initial r_{deg} versus attacked-minus-clean test cross-entropy $M2$ on synthetic hypergraphs. HGCN and HGCN-fixed ($W_e = I$) use $K = 2$, fixed parameters, and incidence attacks with Frobenius norm 5. Results cover four generators, ten replicates, and $\alpha_{deg} \in \{1, 2, 5, 10, 20\}$. Axes are centered within generator-replicate blocks, solid and dashed lines are the corresponding least-squares fits (Appendix F.8).

Figure 2 shows a positive pooled within-block association between initial r_{deg} and $M2$ for both HGCN variants. Within the kNN generator, the association is negative. A plausible explanation is that its degree-control setting also changes hyperedge sizes, overlap, and node co-occurrence patterns, so r_{deg} varies together with other topological properties.

Lemma 1 characterizes how degree imbalance enters the magnitude of operator perturbations. Proposition 1 gives a local replacement-stability lower bound under zero attacks. RQ3 complements these results by examining the empirical association between degree imbalance and attacked test risk across the synthetic generators.

5.4 RQ4: Controlled Depth Dependence

RQ4 isolates the forward depth factor $\rho_{\text{imp}} = v_\sigma C_{\mathbf{P}} C_W$ using fixed propagation operators and scaled-identity layer weights on Cora and CiteSeer. We apply an aligned rank-one feature perturbation δX within ReLU's linear region and measure $M3 = \frac{\|X^{(K)}_{\text{pert}} - X^{(K)}_{\text{clean}}\|_2}{\|\delta X\|_2}$, where $X^{(K)}_{\text{clean}}$ and $X^{(K)}_{\text{pert}}$ are the final-layer representations before and after perturbation, respectively. Across depths $K = 1, \ldots, 6$, $M3$ decreases when $\rho_{\text{imp}} = 0.800$ and increases when $\rho_{\text{imp}} = 1.200$ (Figure 11; Appendix F.9). The observed contraction and expansion follow the geometric forward-depth behavior associated with ρ_{imp}.

6 Conclusion

We analyze adversarial generalization for HGCNs with jointly learned layer matrices and hyperedge weights. On an explicit family and common weight domain, native propagation's optimal global weight sensitivity exceeds that of loopless clique expansion by a factor of order $\sqrt{d_{e,\max}}$. Coupled-SGD structural and feature bounds expose degree, depth, and optimization factors. For fixed architecture, they scale as $\mathcal{O}(T/n)$ under A1–A4, the step-size condition, and loss and gradient envelopes, step sizes, and accumulated expansion bounded independently of n, T. A one-layer, zero-attack construction shows degree-dependent replacement instability over a restricted horizon. RQ1 tracks the directional $\sqrt{m}$ scaling of Corollary 1 on the constructed family, and RQ4 reproduces the geometric forward-depth factor under aligned, untrained conditions. RQ2 shows predominantly decreasing sample-size trends together with several feature-attack reversals, whereas the RQ3 associations vary across generators. Together, these experiments separate controlled tests of the theoretical mechanisms from finite-sample empirical trends.

Reproducibility Statement

Section 3.4 states the assumptions used in the analysis. Appendices C, D, and E provide the technical lemmas, proofs, and supplementary theoretical results. Appendix F specifies the experimental settings, evaluation metrics, and limitations of the archived experimental records. Source code: https://github.com/anonymous/hgcn-adversarial.

AI Use Statement

We use Generative AI tools to polish language, edit LaTeX, and check notation, cross-references, and the presentation and internal consistency of mathematical arguments. Generative AI tools were not used to originate the research questions or core contributions, or to generate experimental code, data, or results. We review all the theoretical claims, proofs, experiments, and reported numerical values. We take full responsibility for the content of the manuscript.

References

Song Bai, Feihu Zhang, and Philip HS Torr. Hypergraph convolution and hypergraph attention. *Pattern Recognition*, 110:107637, 2021.

Aleksandar Bojchevski, Johannes Gasteiger, and Stephan Günnemann. Efficient robustness certificates for discrete data: Sparsity-aware randomized smoothing for graphs, images and more. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119, pp. 1003–1013. PMLR, 2020.

Chang Cao, Han Li, Yulong Wang, Rui Wu, and Hong Chen. Adversarial training for graph convolutional networks: Stability and generalization analysis. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pp. 4797–4805. International Joint Conferences on Artificial Intelligence Organization, 8 2025.

Richard J Chen, Kenneth I Krishkevich, et al. Hypergraph transformer: Weakly-supervised multi-hypergraph learning for whole slide image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Eli Chien, Chao Pan, Jianhao Peng, and Olgica Milenkovic. You are AllSet: A multiset function framework for hypergraph neural networks. In *International Conference on Learning Representations*, 2022.

Farzan Farnia and Asuman Özdağlar. Train simultaneously, generalize better: Stability of gradient-based minimax learners. In *International Conference on Machine Learning*, pp. 3174–3185. PMLR, 2021.

Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. Hypergraph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, number 01, pp. 3558–3565, 2019.

Joel Friedman. A proof of Alon's second eigenvalue conjecture and related problems. *Memoirs of the American Mathematical Society*, 195(910), 2008. doi: 10.1090/memo/0910.

Vikas Garg, Stefanie Jegelka, and Tommi Jaakkola. Generalization and representational limits of graph neural networks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 3419–3430, 2020.

Simon Geisler, Daniel Zügner, and Stephan Günnemann. Reliable graph neural networks via robust aggregation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Simon Geisler, Tobias Schmidt, Hasan Anil Ünal, Daniel Zügner, and Stephan Günnemann. Robustness of graph neural networks at scale. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 7637–7649, 2021.

Moritz Hardt, Benjamin Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pp. 1225–1234, 2016.

Haolin Ju, Dongyue Li, Aneesh Sharma, and Hongyang R. Zhang. Generalization in graph neural networks: Improved pac-bayes bounds on graph diffusion. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 6314–6341. PMLR, 2023.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.

Renjie Liao, Raquel Urtasun, and Richard Zemel. A PAC-Bayesian approach to generalization bounds for graph neural networks. In *International Conference on Learning Representations*, 2021.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=B1QRgziT-.

Jan Schuchardt, Aleksandar Bojchevski, Johannes Gasteiger, and Stephan Günnemann. Collective robustness certificates: Exploiting interdependence in graph neural networks. In *International Conference on Learning Representations*, 2023.

Saurabh Verma and Zhi-Li Zhang. Stability and generalization of graph convolutional neural networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1539–1548, 2019.

Peihao Wang, Shenghao Yang, Yunyu Liu, Zhangyang Wang, and Pan Li. Equivariant hypergraph diffusion neural operators. In *International Conference on Learning Representations (ICLR)*, 2023.

Jiancong Xiao, Yanbo Fan, Ruoyu Sun, Jue Wang, and Zhi-Quan Luo. Stability analysis and generalization bounds of adversarial training. In *Advances in Neural Information Processing Systems*, volume 35, pp. 15446–15459, 2022.

Yue Xing, Qifan Song, and Guang Cheng. On the algorithmic stability of adversarial training. In *Advances in Neural Information Processing Systems*, volume 34, pp. 26523–26535, 2021.

Jiani Zhang et al. Randomized smoothing for certifiable defense against adversarial attacks on hypergraphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Xiang Zhang and Marinka Zitnik. Gnnguard: Defending graph neural networks against adversarial attacks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Dengyong Zhou, Jiayuan Huang, and Bernhard Schölkopf. Learning with hypergraphs: Clustering, classification, and embedding. In *Advances in Neural Information Processing Systems*, volume 19, 2006.

Xianchen Zhou and Hongxia Wang. The generalization error of graph convolutional networks may enlarge with more layers. *Neurocomputing*, 424:97–106, 2021.