STUDYING ADVERSARIAL RESILIENCE OVER GRAPH TOPOLOGY

Anonymous authors
Paper under double-blind review

ABSTRACT

Adversarial attacks to graph analytics are gaining increased attention. To date, two lines of countermeasures have been proposed to resist various graph adversarial attacks from the perspectives of either graph per se or graph neural networks. Nevertheless, a fundamental question lies in whether there exists an intrinsic adversarial resilience state within a graph regime and how to find out such a critical state if exists. This paper contributes to tackle the above research questions from three unique perspectives: i) we regard the process of adversarial learning on graph as a complex multi-object dynamic system, and model the behavior of adversarial attack; ii) we propose a generalized theoretical framework to show the existence of critical adversarial resilience state; and iii) we develop a condensed one-dimensional function to capture the dynamic variation of graph regime under adversarial perturbation, and pinpoint the critical state through solving the equilibrium point of adversarial perturbation-mapping function. Multi-facet experiments are conducted to show that the proposed approach can significantly outperform the state-of-the-art defense methods on five commonly-used real-world datasets.

1 INTRODUCTION

Graph adversarial attacks (GAAs) are launched through executing imperceptive perturbations on edges/weights and even features to the original clean graph (Mu et al., 2021; Li & Wang, 2025), such as injecting/removing edges/nodes, or even modifying features/weights. To date, two main categories of GAAs emerges, one is injection attack (Zheng et al., 2020; Zou et al., 2021; Zhao & Zhang, 2025), i.e. inject new nodes into the original graph rather than revise the existing characteristics. The other is modification attack (Zügner & Günnemann, 2019; Sun et al., 2020; Alom et al., 2025), i.e. directly change the original graph in terms of edges/nodes and/or features/weights.

Accordingly, to defend against various GAAs, a set of adversarial defense mechanisms are proposed from two perspectives: graph per se and graph neural networks (GNNs). For the former, preprocessing countermeasures are utilized, for instance, GCN-SVD (Entezari et al., 2020) utilizes singular value decomposition (SVD) to decompose the adjacency matrix of perturbed graph and obtains a low-rank approximation matrix with the purpose of cleaning adversarial perturbations. GCN-Jaccard (Wu et al., 2019) removes those edges that connected nodes have low Jaccard similarity of features. Analogously, GNNGUARD (Zhang & Zitnik, 2020) assigns higher weights to edges over similar nodes, but prunes edges over dissimilar nodes. Moreover, there also exist optimization countermeasures, e.g., Pro-GNN (Jin et al., 2020) treats adjacency matrix as learning parameters to implement optimization manipulation under the requirements of being close to original adjacency matrix, low-rank of learned adjacency matrix, and feature smoothness. Referring to the variance-based attention, RGCN (Zhu et al., 2019) dispenses differential weights to distinct neighbors during convolution. Additionally, towards the inductive graph-learning tasks, Wang et al. (2023) propose secure graph-learning mechanism in terms of graph partitioning with fairness and balance. Li et al. (2024) propose adversarial training paradigm to resist graph poisoning attack from the out-of-distribution (OOD) perspective. These methods implement adversarial defense by handling graph structure, i.e. graph per se.

On the other hand, the neural ordinary differential equation (ODE) (Chen et al., 2018) and partial differential equation (PDE) (Chamberlain et al., 2021b;a) have been attempted to defend against GAAs. Song et al. (2022) state that graph PDE can validly resist graph topology perturbations by

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.

heat-diffusion model on a general Riemannian manifold. Zhao et al. (2023) propose an energy-conservation graph Hamiltonian flow to enhance adversarial resilience. This line of research tries to adopt diffusion models in physics to conduct graph neural flow w.r.t. the attack resilience, and generates time-evolved embedding vectors for graph nodes, then utilizes backpropagation to minimize loss function for the optimization of model parameters. This category of defense methods focuses on graph neural flow from the diffusion perspective.

Upon the two lines of studies above, we can highlight that: i) GAAs are usually launched through node/edge revision and/or weight/feature modification; ii) towards the graph per se, current adversarial defense methods almost concentrate on low-rank approximation of adjacency matrix of the perturbed graph, deletion of edges over dissimilar nodes, smoothness of node features, or even bias of assigned weights; and iii) towards neural networks, ODE and PDE are employed to infer robust embedding vectors to represent nodes. Our work falls into the category of graph per se, i.e. find out the graph's critical adversarial resilience state through resorting to modifying graph topology only.

We believe the inherent property of graph regime is the most influential factor for model's robustness, only a proper graph topology/structure is built, can the adversarial resilience of graph regime exist. Then, a fundamental research question appears: **whether there exists a critical state of adversarial resilience within each graph regime, if yes, how to find out it.** To surround this fundamental research question, we in theory propose a generalized **Equili**brium point-based adversarial **Res**ilience approach **EquiliRes** to resist various GAAs.

Main Contributions: i) We bridge the connection between adversarial (deep) learning and complex dynamic system (CDS), by which GAAs can be modeled as the process of "oscillation" of CDS; ii) A generalized theoretic framework is proposed to demonstrate the existence of critical stat of adversarial resilience of each graph regime, and explicitly exhibit in what prerequisites and in what perturbation domains such equilibrium point can be accurately located; iii) A condensed one-dimensional function is developed to demonstrate the dynamic variation of graph regime under adversarial perturbations, from which we concretize the equilibrium-point trajectory with asymptotic-stability theory, upon which an attack-resilient graph topology is built accordingly; and iv) Multi-facet experiments are executed to validate the effectiveness of our proposed defense approach on five real-world datasets, and the results show our work remarkably outperforms the state-of-the-art baselines.

2 Theoretical Framework for Graph Adversarial Resilience

2.1 Graph Adversarial Resilience

Prior to bridging the connection between adversarial learning and CDS, we first utilize a simple physical phenomenon to present the intuitive understanding as shown in Fig. 1(a), wherein three balls settle on three positions in a curved orbit. The difference between position P_1 and position P_3 is that the former has no friction while the latter has. At first, the three balls are settled in still, i.e. mathematically $\frac{dp(t)}{dt}\Big|_{t=0}$ =0, it means the first-order derivative of displacement $p(t)$ is zero, that is to say, the balls would keep stationary as the time t changes if no extra action exists, these three positions are what is called equilibrium points. Nevertheless, when a force F acts on the three balls,

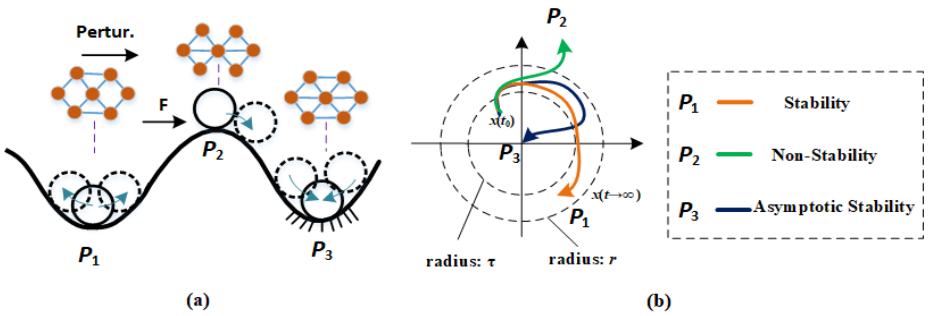


Figure 1: Three balls' motion states and equilibrium points.

for position P_1 the ball would move back-and-forth repeatedly within a bounded range; for position P_2 the ball would move away and cannot go back any more; for position P_3 the ball would first move back-and-forth and finally become still due to the action of friction. Inspired by this physical phenomenon, we ask whether a graph regime also has such equilibrium point to represent the state of adversarial resilience under adversarial perturbations/attacks.

Remark 2.1. Observed from the perspective of complex dynamic systems, the three balls' states respectively correspond to **Stability**, **Non-Stability**, and **Asymptotic Stability** as shown in Fig. 1(b). Stability denotes the ball's trajectory varies along a boundary (e.g. between radius τ and radius r); Non-Stability indicates the ball's trajectory departs from the original position and moves towards infinity (e.g., exceed radius r); Asymptotic Stability implies the ball's trajectory first departs from the original position, and finally stops at the original position (e.g. the origin) even if it may experience a long-time motion. Note that for a dynamic system, there may exist multiple equilibrium points, also called fixed points, but it does not mean all of them are asymptotically stable.

As we know, GAA aims to breach graph's resilience by continuously perturbing graph topology or/and node features. From the viewpoint of CDS, each time the "oscillation" (perturbation) tries to make graph deviate from its original stable state to non-stability state. Hence, we reflectively conjecture that each graph have such an equilibrium point to keep it robust, and detail formally as:

Adversarial Resilience of Graph Regime. For each graph regime, there exists an asymptotically-stable equilibrium point, corresponding to its critical state of adversarial resilience, to keep it attack-resilient under continuous adversarial perturbations, and such a equilibrium point can be acquired as time goes.

GAA generally deceives target/victim model by adding/deleting pivotal nodes/edges, in addition to modifying weights/features. Thus, we can abstractly mirror continuous adversarial perturbations as the dynamic-variation (oscillation) process of complex multi-object physical systems, such as electromagnetic field, charged particles, etc. As analyzed in Fig. 1(b), a dynamic system can fall into the three states: Stability, Non-stability, or Asymptotic Stability. Of course, staring from the stability state, we anticipate the dynamic system can finally converge into the asymptotically-stable state, this implies it can always go back to the original equilibrium point in the end even if a long-time dynamic process undergoes. Thus, referring to the converged trajectory of the ball at position P_3, we define asymptotic stability, and symbol "$\rightharpoonup$" denotes a vector representing the set of objects.

Definition 2.2. (Asymptotic Stability (Khalil, 2015)) Let $\frac{d\vec{x}}{dt} = v(\vec{x})$ be a set of locally Lipschitz functions defined over a domain $\xi \subset \mathbb{R}^n$, and it contains the origin[1], i.e. $v(0)$=0. Then, the equilibrium point x=0 of $\frac{d\vec{x}}{dt}$ is asymptotically stable if there exists $\tau > 0$, s. t. $\left\|\vec{x}\left(t_0\right)\right\| < \tau \Rightarrow \lim_{t\to\infty} \vec{x}\left(t\right) = 0$.

Remark 2.3. For a dynamic system falling in the state of asymptotic stability, it means there exists asymptotically-stable equilibrium point and the system's dynamic variation trajectory will finally converges into such equilibrium point, no matter how long the dynamic may undergo. This process well-matches the behavior of adversarial attacks, that is to say, we can reflectively map the graph adversarial perturbation as dynamic system's "oscillation". In view of this, regarding the adversarial perturbations on graph as dynamic variation of multi-object dynamic system is persuasive.

To make the asymptotically-stable equilibrium point (ASEP) fall into the origin, we give Lyapunov Criterion as stated in Appendix A.

2.2 GAA Modeling

Given graph adversarial attack makes continuous perturbations from two perspectives of single-node autonomy and pairwise-node interaction, especially in present-day deep learning environment,

[1]The origin always can be obtained via coordinate transformation.

thus, graph dynamics (Kundu et al., 2022; Gao et al., 2016) are in general represented as a linear weighted average of the N state variables each of which is associated with a node i, and the nonlinear influence of other nodes on node i. Hence, we can divide the adversarial perturbations over graph into two parts: i) linear perturbation, representing the graph is disturbed by a linear action; and ii) nonlinear perturbation, implying the graph is disturbed by a nonlinear action. As time t goes, the dynamic-variation function under continuous adversarial perturbations can be defined as:

$$\frac{d}{dt}\vec{x}\left(t\right)=A\vec{x}\left(t\right)+B\left(-\vec{\phi}\left(\cdot\right)\right), \tag{1}$$

where $\vec{x}\left(t\right)$ is an N-dimensional vector denoting the states of N nodes, reflecting the linear perturbation, and $\phi(\cdot)$ is a nonlinear-mapping function representing the nonlinear perturbation. Eq. (1) illuminates the dynamic variation of graph nodes under the continuous perturbations (edge removal/injection) by adversarial attack. Accordingly, each node may have a different states at different time t, which is also in concert with the training process of deep learning, namely, during each-round backpropagation-based parameter update, the pre-round output is adopted to compute gradient for the next-round parameter update. In the light of deep-learning principle, we here deem $\phi(\cdot)$ as the input to infer next-round dynamic update, and it should be presented by the output of pre-round perturbation result, in view of this, we can naturally define the output $\vec{y}\left(t\right)=C\vec{x}\left(t\right)$, which represents the pre-round perturbed states of N nodes, thus we have $\phi\left(\cdot\right)=\phi\left(\vec{y}\left(t\right)\right)$ accordingly. Matrices A, B, C are three general parameters, whose elements depend on the particular adversarial attack, the subsequent Eq. (8)–(9) and associated parameters in Table 4 can interpret this point.

The integral from time t_0 to t reflects a period of time aggregation of dynamic variation, which can be equivalently recognized as the process of adversarial learning epoch by epoch, thus the utilization of integral form is adequate to illuminate the effect of continuous adversarial perturbations on graph in a time frame. Therefore, we straightly define the time-aware aggregation of dynamic variation as $\int_{t_0}^{t}x\left(\kappa\right)d\kappa$. Assume a persistent long-time adversarial learning process, i.e. $t\rightarrow\infty$, and simultaneously for the sake of derivative calculation, we introduce *Laplace Transform* for the above time-aware integral of dynamics/perturbation variation

$$L\left[x\left(t\right)\right]=X\left(s\right)=\int_{t_0=0}^{t\rightarrow\infty}x\left(t\right)e^{-st}dt, \tag{2}$$

where $s=\sigma+\omega j, j=\sqrt{-1}$ is a complex number. Given the property of linear transformation of *Laplace Transform*, we have $Y\left(s\right)=C\cdot X\left(s\right)$, and the first derivative's *Laplace Transform* is

$$L\left[\frac{d}{dt}x\left(t\right)\right]=\int_{t_0=0}^{t\rightarrow\infty}\left(\frac{d}{dt}x\left(t\right)\right)e^{-st}dt=\lim_{t\rightarrow\infty}x\left(t\right)e^{-st}-x\left(0\right)+sX\left(s\right)\ . \tag{3}$$

In general, to achieve the convergence, the real part of complex number s must satisfy $\mathrm{Re}\left(s\right)>0$, furthermore, the initial condition meets $x\left(0\right)=0$, thus $L\left[\frac{d}{dt}x\left(t\right)\right]$=$sX\left(s\right)$. Substituting Eq. (2) and Eq. (3) into Eq. (1), an intermediate function $G\left(s\right)$ can be obtained between output and input,

$$\begin{gathered}\vec{X}\left(s\right)=\left(sI-A\right)^{-1}\cdot B\cdot L\left[\vec{\phi}\left(\vec{y}\left(t\right)\right)\right]\\ G\left(s\right)=\frac{\vec{Y}(s)}{L\left[\vec{\phi}\left(\vec{y}\left(t\right)\right)\right]}=\frac{C\cdot\vec{X}(s)}{L\left[\vec{\phi}\left(\vec{y}\left(t\right)\right)\right]}=C\left(sI-A\right)^{-1}B\end{gathered}\ . \tag{4}$$

Remark 2.4. The function $G(s)$ reflects the intermediate variation process of information-passing from an input $L\left[\phi\left(\vec{y}\right)\right]$ to an output $Y(s)$ in *Laplace* form, thus, we can control the coefficients $\{A,B,C\}$ to obtain the result that we anticipate the system to output. From the viewpoint of adversarial resilience, we anticipate the graph finally converges into critical state of adversarial resilience (asymptotically-stable equilibrium point) to remain attack resilient, no mater how long it takes. In fact, you can image function $G(s)$ represents the intermediate process of adversarial learning.

Certainly, we aim to make Eq. (1) converge into the asymptotically-stable equilibrium point, and propose the following theorem on the existence of such equilibrium point.

Theorem 2.5. ***(Existence of Asymptotically-Stable Equilibrium point)*** *The dynamic-variance function (Eq. (1)) has asymptotically-stable equilibrium point(s) if matrix A is Hurwitz, the output $\vec{y}(t)$ and its associated nonlinear-mapping input $\phi\left(\vec{y}(t)\right)$ satisfy the condition that $\vec{\phi}\left(\vec{y}(t)\right)^T \cdot \left[\vec{y}(t) - M\vec{\phi}\left(\vec{y}(t)\right)\right] > 0$, where matrix $M = diag\left({}^1\!/_{k_1}, \cdots, {}^1\!/_{k_p}\right)$, k_i>0. Furthermore, given $\psi = diag(\gamma_1, \cdots, \gamma_p)$, there exists constant $\gamma_i \geq 0$, such that $(1 + \lambda_k\gamma_i) \neq 0$ for each eigenvalue λ_k of matrix A, and meanwhile $M + (I + s\psi)G(s)$ is strictly positive real.*

Proof. See Appendix B

3 Adversarial Resilience Inference

3.1 One-Dimensional Perturbation Mapping

For graph data, the work Kundu et al. (2022) studies both single-node autonomous behavior (SNAB) and pairwise-node interaction behavior (PNIB), taking as a referral in this work, the mentioned adversarial perturbation is investigated along this thought. In detail, we refer to the dynamic variation of graph under adversarial attack as coupled linear function $\chi\left(\vec{x}\right)$ and nonlinear function $\varphi\left(\vec{x}, \vec{x}\right)$. Thus, for node i, we have

$$\frac{dx_i}{dt} = \tilde{A}_{(i)}\chi(x_i) + \sum\nolimits_{j=1}^{N} \tilde{B}_{(ij)}\varphi(x_i, x_j), \tag{5}$$

where $\tilde{A}_{(i)}$ denotes the linear SNAB-caused dynamic-variation coefficient of node i, and $\tilde{B}_{(ij)}$ stands for the nonlinear PNIB-caused dynamic-variation coefficient.

Currently, the node is in general represented in a high-dimension space, to achieve lightweight computation, some dimension-reduction methods are adopted to accomplish low-dimension mapping. Considering the complexity of linkage/interactive relationship over nodes, we here employ one-dimensional condensed method to map the whole graph's dynamic variation. Resorting to Heterogeneous Mean-Field (HMF) approximation theory (Kundu et al., 2022), pursuant to the deduction step by step in Appendix E, we obtain the one-dimensional perturbation-mapping function,

$$\frac{d\tilde{x}}{dt} = \tilde{A}_{(\cdot)}\chi(\tilde{x}) + \tilde{\beta}\varphi(\tilde{x}, \tilde{x}). \tag{6}$$

From Eq. (6), we know the mapping function of the whole graph's dynamics under adversarial perturbation is condensed as a single variable $\tilde{\beta}$, which can also be regarded as the reflection of robustness of the graph as a whole.

Remark 3.1. The condensed one-dimensional mapping function may lead to somewhat loss of graph-structure information. Alternatively, the sophisticated two-dimensional or multi-dimensional mapping functions are workable as well. The principle to develop an appropriate mapping function lies in the capability of faithfully reflecting the caused dynamics/perturbations to the most extent.

3.2 Equilibrium-Point Trajectory

Given that $\chi(x_i)$ is linear and $\varphi(x_i, x_j)$ is nonlinear, we can simply and generally define a dynamic-variation equation to concretize the process of adversarial perturbations as,

$$\frac{dx_i}{dt} = -\tilde{A}_{(i)}x_i + \sum\nolimits_{j=1}^{N} \tilde{B}_{(ij)}\frac{Hx_j^2}{\theta x_j^2 + 1}. \tag{7}$$

The first term on the right hand of Eq. (7) presents node i's self-dynamics, for detail, we utilize the degree centrality of node i to represent the probability of being attacked. The second stands for the interactively-affected perturbation by its neighboring nodes. This simple yet effective one-dimensional function is developed to map the dynamics of graph regime as a whole,

in fact, two-dimensional mapping or multi-dimensional mapping can be explored as well for distinct graph topologies, since our proposed asymptotically-stable theoretic framework fully supports multi-dimensional dynamic-variation function. The more accurate the modeling on GAAs, the higher the adversarial resilience would be. To get an accurate function to capture the dynamics, usually, some extra variable $\theta > 0$ is needed, and H denotes the average in-degree weight of node i, implying the affected extent of perturbation by neighbors. Referring to Eq. (6) in the condensed one-dimensional form, we can concretize Eq. (7) as

$$\frac{d\tilde{x}}{dt} = -a\tilde{x} + \tilde{\beta}\frac{H\tilde{x}^2}{\theta\tilde{x}^2 + 1}, \tag{8}$$

where a is a successor constant condensed from vector $\tilde{A}_{(i)}$. The equation above exhibits the detailed settings regrading the concrete graph attributes: degree centrality, in-degree weight, etc. Here what we need emphasize is that Eq. (6) is a theoretical guidance, Eq. (7) and Eq. (8) are the concretized deductions w.r.t. specified graph topology, which implies different graph regimes may have distinct formulae. From Eq. (8), we can infer the following theorems.

Theorem 3.2. *Upon the condensed one-dimensional mapping function, a graph regime disturbed by graph adversarial attack can be finally converged into an asymptotically-stable equilibrium point if the constant k in Theorem 2.5 belongs to the interval $\left[H \cdot \left(2\sqrt{\theta}\right)^{-1}, \infty\right)$.*

Proof. See Appendix F.

Theorem 3.2 evidence the adversarial perturbation domain, only within which can the asymptotically-stable equilibrium point be finally located, and the following Theorem 3.3 warrants the existence of such equilibrium point.

Theorem 3.3. *The condensed one-dimensional dynamic-variation function in Eq. (8) is asymptotically stable.*

Proof. See Appendix G.

Remark 3.4. Theorem 3.2 and Theorem 3.3 guarantee our proposed one-dimensional dynamic-variation function has asymptotically-stable equilibrium point, i.e. there exist a critical state of adversarial resilience to keep graph regime attack-resilient. The parameter $\omega \in \mathbb{R}$ in proof of Theorem 3.3 means the equilibrium point can be found out in the real number field, simultaneously with the constraint of perturbation scope of k. In other words, the existence of equilibrium point can be only warranted within a bounded domain. Once the perturbation jumps out of the constraint domain, the critical state of adversarial resilience would disappear.

Seen from Eq. (8), we know $\frac{d\tilde{x}}{dt}$=0 if $\tilde{x}$ = 0, which means the entire graph is completely broken with no connection between nodes (Eq. (33)), this is also a so-called equilibrium point, but not what we anticipate although this case indeed happens in reality. Thereby, we need calculate other equilibrium points divided by $\tilde{x}$, and at the same time regard $\tilde{x}$ as the variable of $\tilde{\beta}$,

$$\tilde{\beta}\left(\tilde{x}\right) = a \cdot \frac{\theta\tilde{x}^2 + 1}{H\tilde{x}}. \tag{9}$$

Pursuant to the above function, we know that a trajectory consisting of infinite equilibrium points can be drawn in theory. Since our work focuses on enhancing adversarial resilience only from the angle of graph topology, i.e. resist GAAs from the radical graph structure. In order to endow appropriate variables for Eq. (8), we choose the basic graph property–Degree Centrality, which reflects the "hub" role in the connectivity of entire graph, and defined as $C_{deg}(i) = {|d_i|}/{(N-1)}$ where d_i denotes the set of neighbors of node i. Accordingly, we can separately obtain in-degree centrality $C_{Indeg}(i) = {|d_{Inneib}(i)|}/{(N-1)}$ where $d_{Inneib}(i)$ denotes the set of node i's in-going neighbors, and out-degree centrality $C_{Outdeg}(i) = {|d_{Outneib}(i)|}/{(N-1)}$ where $d_{Outneib}(i)$ indicates the set of node i's out-going neighbors. As mentioned previously, we set $a = \eta \cdot \frac{\sum_i C_{Indeg}(i) + \sum_i C_{Outdeg}(i)}{N}$,

$H = \sum_i d_{Inneib}(i)/_N$, and use the deduced Eq. (33) in Appendix E to calculate $\tilde{x}$. Although the degree centrality is the basic and simple attribute in graph regime, The subsequent experiment showcases this basic and simple degree-related property plays an important role to boost adversarial resilience under the guidance of equilibrium-point trajectory. that is to say, the attack-resilient graph topology can be established resorting to this basic degree-related properties.

3.3 Quantifying Perturbation Bound

Theoretically, the perturbed-bound with attack budget can be quantified as sketched in Appendix H.

4 Implementation

The asymptotically-stable equilibrium point is in accordance with the critical state of adversarial resilience for graph regime, thus we take such an equilibrium point as the referral to build a new attack-resilient graph topology for perturbed graph. In detail, resorting to the adjacency matrix of undisturbed graph, we first compute $\tilde{\beta}$ using Eq. (9) to obtain an asymptotically-stable equilibrium point as shown "Original" in Fig. 2–6 in Appendix J.2. Section 5.2 will detail the computation of equilibrium-point trajectory. Then, we set up an iterative-optimization procedure to generate a robust adjacency matrix referring to this asymptotically-stable equilibrium point, i.e., enforcing the modified adjacency matrix-based coordinator point to be close to the "Original" point. For easy understanding, we state the procedures in Algorithm 1 and Algorithm 2 in Appendix I.

5 Experiment Evaluation

Besides performance analytics, our experiments need to answer three fundamental research questions: i) **RQ1:** Whether there exists a potential correlation between adversarial learning and CDSs indeed; ii) **RQ2:** Whether *Laplace Transfer* is adaptable to embody the process of continuous adversarial perturbations; and iii) **RQ3:** Whether the one-dimensional mapping function suffices to appropriately present the perturbation-caused dynamics over graph.

5.1 Configuration

The statistics of five commonly-used moderate- and large-scale datasets[2]: Polblogs, Cora, Cora_ML, Citeseer, and Amazon Photo, seven representative adversarial defense baselines (GCN (Kipf & Welling, 2017), GAT (Velickovic et al., 2018), RGCN (Zhu et al., 2019), GCN-SVD (Entezari et al., 2020), HANG (Zhao et al., 2023), Mid-GCN (Huang et al., 2023), and GOOD-AT (Li et al., 2024)), three non-targeted attacks (Metattack (Zügner & Günnemann, 2019), CE-PGD (Xu et al., 2019), DICE (Waniek et al., 2016)), one adaptive attack, and the detailed execution settings are stated in Appendix J.1.

5.2 Equilibrium-Point Trajectory

To predict the adverseness of graph adversarial attacks, we use the parameters in Table 4 in Appendix J.2 to draw equilibrium-point trajectories with Polblogs, Amazon Photo, Cora_ML, Cora, and Citeseer in Figs. 2-6, from which we observe the adversarial resilience degrades almost along the equilibrium-point trajectory curves, this phenomenon unveils three-aspect hints: i) the concordance of equilibrium-point trajectory and adverseness of adversarial attacks discloses the existence of correlation between adversarial learning and complex dynamic system, which answers the first question **RQ1**; ii) note that for each three curves drawn from the five datasets, the variation tendency of inferred adversarial resilience under different-level perturbation strength ranging from 3% to 30% goes exactly along the theoretical equilibrium-point curve which is deduced by *Laplace Transfer*, this unveils the rationality of introduction of *Laplace Transfer* for the modeling on graph adversarial attacks, which replies the second question **RQ2**; and iii) the condensed one-dimensional function-based equilibrium-point trajectory is appropriate to capture the dynamics of graph regime as a whole under the continuous adversarial perturbations, which answers the third question **RQ3**.

[2]https://pytorch-geometric.readthedocs.io/en/latest/cheatsheet/data_cheatsheet.html

Table 1: Accuracy (%) of node classification on Polblogs (**bold**-the best)

GAA	Def. \ Per.	0%	5%	10%	15%	20%	25%
Metattack	GCN	94.26±1.24	76.35±1.39	70.61±1.51	67.66±1.39	67.45±1.45	66.34±1.54
	GCN-SVD	93.98±0.52	**92.52±1.09**	87.99±4.35	71.13±2.93	67.25±1.69	65.56±1.49
	EquiliRes	**94.34±0.81** (**0.08%**↑)	90.11±2.62 (**2.41%**↓)	**89.70±2.22** (**1.71%**↑)	**86.53±1.47** (**15.40%**↑)	**81.41±1.94** (**13.96%**↑)	**73.11±2.60** (**6.97%**↑)
CE-PGD	GCN	94.26±1.24	84.07±0.65	79.19±1.08	76.86±0.91	75.37±1.05	74.47±1.14
	GCN-SVD	93.98±0.52	85.31±1.06	79.43±1.04	76.82±1.06	75.59±1.00	74.80±1.26
	EquiliRes	**94.34±0.81** (**0.08%**↑)	**86.69±1.19** (**1.38%**↑)	**82.67±1.12** (**3.24%**↑)	**79.93±1.39** (**3.07%**↑)	**76.82±1.01** (**1.23%**↑)	**75.59±1.41** (**0.79%**↑)
DICE	GCN	94.26±1.24	86.58±1.22	80.83±0.96	77.42±1.07	75.44±1.13	73.37±0.99
	GCN-SVD	93.98±0.52	**90.83±0.78**	88.29±0.50	85.56±0.66	82.92±1.20	81.29±0.96
	EquiliRes	**94.34±0.81** (**0.08%**↑)	89.47±0.95 (**1.36%**↓)	**88.43±2.32** (**0.14%**↑)	**86.85±2.11** (**1.29%**↑)	**84.75±1.54** (**1.83%**↑)	**83.45±1.63** (**2.16%**↑)

5.3 Performance Analytics

5.3.1 Adversarial Resilience Promotion under GAAs

Asymptotically-Stable Equilibrium-Point Trajectory Conducts Attack-Resilient Topology. The experimental results in Table 1, Tables 5–6 in Appendix J.3 showcase the remarkable robustness improvement of our EquiliRes compared to GCN-SVD and GCN. The adversarial perturbation rate (APR) 0% means there does not exist adversarial attack, we execute matrix-modification operation with a little percentage 1%, from which, we can deduce two-facet opinions: i) the accuracy of node classification is high without adversarial perturbation, e.g. approximately 94% with the three methods on Polblogs dataset; and ii) our EquiliRes gains larger accuracy compared to the two baselines, which implies our proposed asymptotic stability theory can validly locate the critical state of adversarial resilience to keep robust. Even for clean graphs, our method still can further boost them.

For Metattack attack, the accuracy raises by 15.40%, 13.96%, 6.97% than the second best when the APRs are 15%, 20% and 25% on Polblogs, and by 5.67%, 6.51% and 4.93% on Citeseer when APRs are 10%, 15% and 20% respectively, which shows our proposed approach can effectively defend against such non-targeted gradient-based attack. For CE-PGD attack, although it perturbs clean graph in a maximum probability through enlarging the cross-entropy loss, pursuant to Theorem 3.3, our proposed dynamic-variation function can finally ends up at asymptotically-stable equilibrium point to make the graph attack-resilient, thus it gains 3.24%, 3.07% and 1.23% improvement on Polblogs, and 2.65%, 2.53% and 2.30% on Citeseer when APRs are 10%, 15% and 20% respectively.

Taking Equilibrium State of Graph Regime as Booster to Promote Adversarial Resilience. To furtherly explore the effectiveness, we at first find out the ASEP and generate associated topology, then utilize the existing methods to execute node-classification task. Table 2 and Table 7 (Appendix J.3) state the results, wherein the sign "**+**" indicates our EquiliRes as the preliminary is first executed, then followed by GAT, RGCN, HANG, Mid-GCN, and GOOD-AT.

EquiliRes as a booster can enhance the existing methods, upon which the different-level advancement on the five datasets at APR 0% unveils that even a clean graph can be enhanced by constructing an equilibrium-state topology at first. Moreover, GAT improves the accuracy by on-average 11.90% on Polblogs, 10.916% on Amazon Photo, 4.68% on Cora_ML, 2.02% on Cora, 2.58% on Citeseer, HANG raises by on-average 1.70%, 0.82%, 5.02%, 1.99%, 2.55%, Mid-GCN promotes by on-average 2.07%, 1.18%, 0.42%, 2.20%, 0.61%, as well as RGCN and GOOD-AT with enhancement.

5.3.2 Performance under Adaptive Attack

We also introduce Adaptive Attack w.r.t. relaxing the adjacency to the interval $[0, 1]^{n\times n}$ instead of the original binary element (0 or 1), seen from the results reported in Appendix J.4, we can observe that although the performance under Adaptive Attack is remarkably degraded compared

Table 2: Accuracy (%) of node classification under combination with EquiliRes using Metattack

Dataset	Def. \ Per.	0%	5%	10%	15%	20%	25%
Polblogs	GAT	94.88±0.42	86.93±2.39	76.89±2.14	71.78±1.46	68.29±1.83	66.99±1.64
	Ours+GAT	**95.12±0.52** (**0.24%**↑)	**92.21±1.20** (**5.28%**↑)	**89.94±1.89** (**13.05%**↑)	**87.96±2.09** (**16.18%**↑)	**84.17±2.26** (**15.88%**↑)	**76.08±4.80** (**9.09%**↑)
	RGCN	94.87±0.22	77.87±0.69	70.50±0.37	67.60±0.24	68.35±0.18	67.68±0.30
	Ours+RGCN	**95.15±0.28** (**0.28%**↑)	**84.16±0.33** (**6.29%**↑)	**77.37±0.42** (**6.87%**↑)	**75.03±0.31** (**7.43%**↑)	**72.42±0.28** (**4.07%**↑)	**70.05±0.25** (**2.37%**↑)
	HANG	93.72±1.26	88.88±1.83	87.45±1.21	86.53±1.47	81.41±1.94	73.11±2.60
	Ours+HANG	**93.85±0.75** (**0.13%**↑)	**90.73±1.14** (**1.85%**↑)	**89.65±2.13** (**2.20%**↑)	**86.67±2.08** (**0.14%**↑)	**82.71±2.66** (**1.30%**↑)	**76.11±4.53** (**3.00%**↑)
	Mid-GCN	91.11±1.82	89.44±3.35	84.71±4.94	72.59±4.10	67.16±1.38	64.70±1.08
	Ours+Mid-GCN	**92.15±2.26** (**1.04%**↑)	**90.82±2.49** (**1.38%**↑)	**85.27±4.67** (**0.56%**↑)	**74.29±1.96** (**1.70%**↑)	**70.88±1.51** (**3.72%**↑)	**67.67±1.39** (**2.97%**↑)
	GOOD-AT	95.17±0.21	94.78±0.52	94.97±0.34	93.50±0.98	93.68±1.36	93.68±1.36
	Ours+GOOD-AT	**95.27±0.20** (**0.10%**↑)	**95.42±0.27** (**0.64%**↑)	**95.17±0.47** (**0.20%**↑)	**95.24±0.35** (**1.74%**↑)	**94.05±0.57** (**0.37%**↑)	**93.80±0.96** (**0.12%**↑)

to that under other adversarial attacks like Metattack in Table 2, our EquiliRes still can retain a comparative performance. This also reflects that the designed decimal-based Adaptive Attack is inapplicable for GAT and Mid-GCN, but applicable for our EquiliRes and HANG.

5.3.3 Ablation Studies

We also perform ablation studies in the perspectives of with/without ASEP, different-level attack budget, from the results in Appendix J.5, we can observe the performance with ASEP as a guidance significantly outperforms that without it. Furthermore, the derived equilibrium points $(\tilde{\beta}, \tilde{x})$ under different-level attack budgets can appropriately locate at the theoretical ASEP trajectory.

5.3.4 Variation of Rank/Singular Values and Computational Overhead

GAA could enlarge the rank and singular values of adjacency matrix (Jin et al., 2020; Zhao et al., 2023), to resist that, our experiments showcase EquiliRes can indeed decease them when the graph topology matches the ASEP state. The analytics are detailed in Appendices J.6 and J.7. The analysis on the computation complexity and its scalability is additionally detailed in Appendix J.8.

6 Related Work

We review the related work in Appendix K. Simultaneously, we also discuss the behind motivation why to bridge the connection between adversarial learning and CDSs, along with three open issues from the perspectives of condensed one-dimensional mapping function, adversarial resilience on multi-modal data, and parameters of resilience-declining function in Appendix L.

7 Conclusion

Regarding the difficulty of modeling the perturbations of graph regime under various adversarial attacks, we attempt to study the intrinsic adversarial mechanism resorting to asymptotic-stability theory, and construct a generalized theoretical framework for critical state of adversarial resilience, by which we can find out the ASEP of the perturbation-mapping function to represent the graph's critical state of adversarial resilience. Multi-facet experiments validate the effectiveness and the rationality of bridging the connection between adversarial perturbation in deep learning and "oscillation" in complex dynamic systems. To us, the most important innovation is to contribute a provable and interpretable solution to pinpoint the critical state of adversarial resilience for graph data. We believe our work will stimulate more explorations on the intrinsic adversarial resilience in future.

AI USE STATEMENT

We did not use AI tools.

ETHICS STATEMENT

In developing our EquiliRes adversarial defense approach, we have carefully considered the ethical implications and potential security problem of our research. This research mainly proposes a generalized theory framework for critical state of adversarial resilience and concrete adversarial defense algorithm for graph regime. Of course, in order to validate the effectiveness and rationality of our proposed theoretic framework, we also introduce three existing non-targeted graph adversarial attacks and one specifically-designed adaptive attack for our approach, which may potential bring security risks to the public community. Nevertheless, we have offer several countermeasures and assistance resources to ensure that our study are conducted ethically and responsibly.

Stakeholders and Potential Impact. The key stakeholders engaged in our research include data owner, adversarial attacker and defender, and the wider public who might be potentially affected by misuse of these attack techniques. First, consider from the angle of personal privacy, according to the descriptions and statistics of the five used datasets in this paper, they are all open-sourced datasets and not straightly related to personal sensitive information and privacy. For example, the datasets Cora_ML, Cora are scientific publication-citation networks, the citation links between scientific publications reflect the academic correlation between/within different research subjects/directions, without the disclosure of personal privacy. Second, even some graph adversarial attacks are misused, we also introduce some more advanced mitigation to resist these attacks, also our benchmark including the source codes for the mentioned adversarial attacks and adversarial defense methods has been released at the anonymous website during the peer-review phase. According the research results, the defenders can enhance the resilience/robustness of graph/network-focused systems or platforms using our open-soured codes. Third, to mitigate these security risks, we also comprehensively analyze several representative graph adversarial attacks and multiple state-of-the-art adversarial defense mechanisms, especially the analytics on our proposed EquiliRes from the viewpoint of both theory and interpretability, in addition to emphasizing the responsible usage and development of the mentioned adversarial attacks and mitigations, explicitly discourage any unethical applications.

Responsible Disclosure and Concerns. Graph adversarial attacks (GAAs) and specifically-designed adaptive attack will bring security risk for graph/network-based systems or platforms, thus, the GAAs and adaptive attack may potentially cause severe ethical challenges to many graph-analytics domains/tasks, ranging from social network analysis and entity mining, chemistry (molecule structure) investigation, finance (trading network) anti-fraud and risk control, biomedical innovation, etc. To address these risks and concerns, we establish a restricted experimental environment to perform GAAs and adversarial defense studies, only allowing the research team to access the benchmark in the process of analyzing the severe vulnerabilities of defense approaches, which minimizes the risks of disclosing the inherent vulnerabilities of existing works, regarding the potential misuse by attackers. In addition, our institute and universities also provide psychological services for the research member all the time.

Balancing Potential Benefits and Harms. In recent years, the imperceptible adversarial perturbations by GAAs could make the deep learning models infer totally wrong results, which would bring serious consequence to many realistic tasks. For example, the deep learning models are employed to predict new molecule structure in chemistry and biopharmaceutics, the incorrect prediction output may yield an unfunctional chemical compound and medicine, which may cause potential danger and threat to human life, thereby, it becomes an extremely important issue on how to design a robust and trustworthy deep learning model to enhance the recognition capability even some perturbations exist. To achieve this goal, our work contributes a resilience-enhancing mechanism from the standpoint of asymptotic-stability theory derived from complex dynamic systems, by which the critical state of adversarial resilience can be located appropriately for each graph regime. In view of this, our proposed method can assist to find out the intrinsic robust graph topology. Therefore, we believe our work's research value outweighs the potential harms, even though such critical state of graph regime can be manipulated by attacker.

REPRODUCIBILITY STATEMENT

In our work, all the experiment results of our approach and other baselines are performed in an open-source benchmark for graph adversarial attacks and adversarial defense mechanism, thus, all the experimental results are reproducible. We publicize our source code at Github to enhance the sharing on our scientific findings.

Data Collection and Use.

The five datasets used in our experiments, i.e. Polblogs, Cora_ML, Cora, Citeseer, and Amazon Photo are all open-sourced and commonly-used in academic domain. They all can be downloaded directly from Internet:`https://pytorch-geometric.readthedocs.io/en/latest/cheatsheet/data_cheatsheet.html`.

Source Code.

We open-source the code at Github: `https://github.com/Anony4OpenScience/EquiliRes_V1.0`.

ACKNOWLEDGMENTS

We thank our collaborators for their valuable discussions, suggestions, and feedback. We also acknowledge the developers of the open-source models, datasets that support the baseline experiments and data analysis.

REFERENCES

Zulfikar Alom, Tran Gia Bao Ngo, Murat Kantarcioglu, and Cuneyt Gurcan Akcora. Gottack: Universal adversarial attacks on graph neural networks via graph orbits learning. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28*, 2025.

Aleksandar Bojchevski and Stephan Günnemann. Adversarial attacks on node embeddings via graph poisoning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 695–704, 2019.

Ben Chamberlain, James Rowbottom, Davide Eynard, Francesco Di Giovanni, Xiaowen Dong, and Michael M. Bronstein. Beltrami flow and neural diffusion on graphs. In *34th Conference on Neural Information Processing Systems*, pp. 1594–1609, 2021a.

Ben Chamberlain, James Rowbottom, Maria I. Gorinova, Michael M. Bronstein, Stefan Webb, and Emanuele Rossi. GRAND: graph neural diffusion. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 1407–1418, 2021b.

Chi-Tsong Chen. *Linear System Theory and Design*. Oxford University Press, New York, third edition edition, 1999. ISBN 0-19-511777-8.

TianQi Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In *31th Conference on Neural Information Processing Systems*, pp. 6572–6583, 2018.

Authors: R. J. Duffin and D. Hazony. The degree of a rational matrix function. *Journal of the Society for Industrial and Applied Mathematics*, 11(3):645–658, 1963.

Negin Entezari, Saba A. Al-Sayouri, Amirali Darvishzadeh, and Evangelos E. Papalexakis. All you need is low (rank): Defending against adversarial attacks on graphs. In *The Thirteenth ACM International Conference on Web Search and Data Mining (WSDM '20)*, pp. 169–177, 2020.

Jianxi Gao, Baruch Barzel, and Albert-László Barabási. Universal resilience patterns in complex networks. *Nature*, 530:307–312, 2016.

B. L. Ho and R. E. Kalman. Effective construction of linear state-variable models from input/output functions. *AUTOMATISIERUNGS-TECHNIK*, 14(12):545–548, 1966.

Jincheng Huang, Lun Du, Xu Chen, Qiang Fu, Shi Han, and Dongmei Zhang. Robust mid-pass filtering graph convolutional networks. In *Proceedings of the ACM Web Conference 2023*, pp. 328–338, 2023.

Wei Jin, Yao Ma, Xiaorui Liu, Xianfeng Tang, Suhang Wang, and Jiliang Tang. Graph structure learning for robust graph neural networks. In *The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pp. 66–74, 2020.

Hassank K. Khalil. *Nonlinear Systems*. Prentice Hall Prentice-Hall, Inc., third edition edition, 2002.

Hassank K. Khalil. *Nonlinear Control*. Pearson Education, Inc., 2015.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations*, 2017.

Prosenjit Kundu, Hiroshi Kori, and Naoki Masuda. Accuracy of a one-dimensional reduction of dynamical systems on networks. *PHYSICAL REVIEW E*, 105(2):024305(14), 2022.

Jiate Li and Binghui Wang. AGNNCert: Defending graph neural networks against arbitrary perturbations with deterministic certification. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 7251–7270, 2025.

Kuan Li, Yiwen Chen, Yang Liu, Jin Wang, Qing He, Minhao Cheng, and Xiang Ao. Boosting the adversarial robustness of graph neural networks: An OOD perspective. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11*, 2024.

Xuanqing Liu, Si Si, Jerry Zhu, Yang Li, and Cho-Jui Hsieh. A unified framework for data poisoning attack to graph-based semi-supervised learning. In *32th Conference on Neural Information Processing Systems*, pp. 9777–9787, 2019.

Jiaming Mu, Binghui Wang, Qi Li, Kun Sun, Mingwei Xu, and Zhuotao Liu. A hard label black-box adversarial attack against graph neural networks. In *2021 ACM SIGSAC Conference on Computer and Communications Security*, pp. 108–125, 2021.

T. Konstantin Rusch, Ben Chamberlain, James Rowbottom, Siddhartha Mishra, and Michael M. Bronstein. Graph-coupled oscillator networks. In *International Conference on Machine Learning*, volume 162, pp. 18888–18909, 2022.

Yang Song, Qiyu Kang, Sijie Wang, Kai Zhao, and Wee Peng Tay. On the robustness of graph neural diffusion to topology perturbations. In *35th Conference on Neural Information Processing Systems*, 2022.

Yiwei Sun, Suhang Wang, Xianfeng Tang, Tsung-Yu Hsieh, and Vasant G. Honavar. Adversarial attacks on graph neural networks via node injections: A hierarchical reinforcement learning approach. In *The Web Conference 2020*, pp. 673–683, 2020.

Matthew Thorpe, Tan Minh Nguyen, Hedi Xia, Thomas Strohmer, Andrea L. Bertozzi, Stanley J. Osher, and Bao Wang. GRAND++: graph neural diffusion with A source term. In *The Tenth International Conference on Learning Representations*, 2022.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *6th International Conference on Learning Representations*, 2018.

Binghui Wang and Neil Zhenqiang Gong. Attacking graph-based classification via manipulating the graph structure. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pp. 2023 – 2040, 2019.

Cheng-Long Wang, Mengdi Huai, and Di Wang. Inductive graph unlearning. In *32nd USENIX Security Symposium*, pp. 3205–3222, 2023.

Marcin Waniek, Tomasz P. Michalak, Talal Rahwan, and Michael J. Wooldridge. Hiding individuals and communities in a social network. *arXiv*, abs/1608.00375, 2016.

Huijun Wu, Chen Wang, Yuriy Tyshetskiy, Andrew Docherty, Kai Lu, and Liming Zhu. Adversarial examples for graph data: Deep insights into attack and defense. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pp. 4816–4823, 2019.

Kaidi Xu, Hongge Chen, Sijia Liu, Pin-Yu Chen, Tsui-Wei Weng, Mingyi Hong, and Xue Lin. Topology attack and defense for graph neural networks: An optimization perspective. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pp. 3961–3967, 2019.

Dante C. Youla. On the factorization of rational matrices. *IRE Trans. Inf. Theory*, 7(3):172–189, 1961.

Xiang Zhang and Marinka Zitnik. Gnnguard: Defending graph neural networks against adversarial attacks. In *33th Conference on Neural Information Processing Systems*, 2020.

Kai Zhao, Qiyu Kang, Yang Song, Rui She, Sijie Wang, and Wee Peng Tay. Adversarial robustness in graph neural networks: A hamiltonian approach. In *36th Conference on Neural Information Processing Systems*, 2023.

Maochang Zhao and Jing Zhang. Highly imperceptible black-box graph injection attacks with reinforcement learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(12): 13357–13364, 2025.

Qinkai Zheng, Yixiao Fei, Yanhao Li, Qingmin Liu, Minhao Hu, and Qibo Sun. Kdd cup 2020 ml track 2 adversarial attacks and defense on academic graph 1st place solution. In *KDD cup 2020 ML Track 2*, 2020.

Dingyuan Zhu, Ziwei Zhang, Peng Cui, and Wenwu Zhu. Robust graph convolutional networks against adversarial attacks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1399–1407, 2019.

Xu Zou, Qinkai Zheng, Yuxiao Dong, Xinyu Guan, Evgeny Kharlamov, Jialiang Lu, and Jie Tang. TDGIA: effective injection attacks on graph neural networks. In *The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2461–2471, 2021.

Daniel Zügner and Stephan Günnemann. Adversarial attacks on graph neural networks via meta learning. In *7th International Conference on Learning Representations*, 2019.

A LYAPUNOV CRITERION

In the domain of system stability, Lyapunov criterion as a classic methodology provides the strict condition to guarantee the origin must be an asymptotically-stable equilibrium point for multi-object dynamic systems. For the detailed proof, refer to the reference Khalil (2015).

Theorem A.1. ***(Lyapunov Criterion (Khalil, 2015))*** *Let $v(\vec{x})$ be a locally Lipschitz function over a domain $\xi \subset \mathbb{R}^n$, which contains the origin. Define a continuously differentiable function $V\left(\vec{x}\right)$ over domain ξ, such that*

$$\begin{aligned} V\left(0\right) = 0, V\left(\vec{x}\right) > 0 \;\; for\; all \;\; \vec{x} \in \xi \;\; and \;\; \vec{x} \neq 0 \\ \frac{d}{dt} V\left(\vec{x}\right) < 0 \;\; for \;\; all \;\; \vec{x} \in \xi \;\; and \;\; \vec{x} \neq 0 \end{aligned} . \tag{10}$$

This Lyapunov criterion gives a general methodology to prove whether an asymptotically-stable equilibrium point exists or not, upon which, we infer the subsequent theorems in this paper.

B PROOF OF THEOREM 2.5

Pursuant to Theorem A.1, a candidate Lyapunov function is defined at first, then the quadratic-form is employed for the linear-part perturbation. For the conciseness, variable t is omitted hereafter,

$$V_{Linear}\left(\vec{x}\right) = \frac{1}{2}\vec{x}^T P\vec{x}. \tag{11}$$

For the nonlinear-part perturbation, a reasonable and valid manner is to use the integral form to present the aggregated effect of continuous perturbations,

$$V_{Nonlinear}\left(\vec{x}\right) = \sum\nolimits_{i=1}^{m} \gamma_i \int_0^{y_i} \phi_i(\sigma)d\sigma. \tag{12}$$

According to Theorem A.1 and Eq. (1), we know V_{Linear} ought to be positive definite, this implies the matrix must be $P > 0$ and symmetric. The first-order derivative of linear-part perturbation and nonlinear-part perturbation can be respectively calculated as:

$$\frac{d}{dt}V_{Linear}\left(\vec{x}\right) = \frac{1}{2}\vec{x}^T\left(PA + A^T P\right)\vec{x} + \vec{x}^T PB\left(-\phi\left(\vec{y}\right)\right). \tag{13}$$

$$\frac{d}{dt}V_{Nonlinear}\left(\vec{x}\right) = \phi^T\left(\vec{y}\right)\psi C\left(A\vec{x} + B\left(-\phi\left(\vec{y}\right)\right)\right). \tag{14}$$

Referring to the aforementioned Lyapunov Criterion, $(PA + A^T P)$ is anticipated to be negative definite, to prove this point, the following theorem is provided.

Theorem B.1. *For any given positive definite and symmetric matrix $\tilde{N}$, and its decomposed non-singular matrix $\hat{N}$, $PA + A^T P = -\hat{N}^T\hat{N}$ has a unique positive definite and symmetric solution P if all eigenvalues of matrix A have negative-real parts.*

Proof. See Appendix C.

Theorem B.1 guarantees the negative definite of first part of Eq. (13). Next, turn to the rest part $\vec{x}^T PB\left(-\phi\left(\vec{y}\right)\right)$ and $\frac{d}{dt}V_{Nonlinear}\left(\vec{x}\right)$. Similarly, we also attempt to match a quadratic form,

$$\begin{aligned}\frac{d}{dt}V\left(\vec{x}\right) &= \frac{1}{2}\vec{x}^T\left(-\hat{N}^T\hat{N}\right)\vec{x} + \vec{x}^T PB\left(-\phi\left(\vec{y}\right)\right) \\ &\quad + \phi\left(\vec{y}\right)^T \psi C\left(A\vec{x} + B\left(-\phi\left(\vec{y}\right)\right)\right) \\ &= -\frac{1}{2}\left(\hat{N}\vec{x} - W\phi\left(\vec{y}\right)\right)^T\left(\hat{N}\vec{x} - W\phi\left(\vec{y}\right)\right) \\ &\quad - \phi\left(\vec{y}\right)^T\left[\vec{y} - M\phi\left(\vec{y}\right)\right]\end{aligned}. \tag{15}$$

On account of the negative semidefinite of $\frac{d}{dt}V\left(\vec{x}\right)$, we deduce the following theorem.

Theorem B.2. *There exist a diagonal matrix M, such that the perturbation output $\vec{y}$ and its associated nonlinear-mapping input $\phi\left(\vec{y}\right)$ satisfying $\vec{\phi}\left(y\right)^T \cdot \left[\vec{y} - M\vec{\phi}\left(y\right)\right] \geq 0$, a matrix W and a diagonal matrix ψ, such that $PB = -\hat{N}^T W + A^T C^T\psi + C^T$, $W^T W = \psi CB + 2M + B^T C^T\psi$.*

Proof. See Appendix D.

Theorem B.1 and Theorem B.2 warrant the realization of Eq. (15), namely the first-order derivative of candidate Lyapunov function is negative semidefinite. To finally warrant it converges into the asymptotically-stable equilibrium point, we anticipate it to be negative definite categorically,

$$\frac{d}{dt}V\left(\vec{x}\right) < -\frac{1}{2}\varepsilon\vec{x}^T P\vec{x}, \tag{16}$$

where ε is an arbitrarily small positive number. To achieve this requirement, Eq. (15) must satisfy $PA + A^T P = -\hat{N}^T\hat{N} - \varepsilon P$. Now, we need prove the existence of such an ε. According to

Theorem B.1, we know if $(PA + A^T P)$ is positive definite and symmetric, then it can be equal to the multiplication of a nonsingular matrix by its transpose. That is to say, we need prove $\hat{N}^T\hat{N} + \varepsilon P$ is positive definite (obviously symmetric), thus we have

$$\begin{aligned} &\vec{x}^T\hat{N}^T\hat{N}\vec{x} + \varepsilon\int_0^t \vec{x}^T e^{\left(A^T\sigma\right)}\hat{N}^T\hat{N}e^{(A\sigma)}\vec{x}d\sigma > 0 \\ &\varepsilon > -\frac{\left\|\vec{x}\hat{N}\right\|_2^2}{\int_{t_0}^t \left\|\vec{x}e^{(A\sigma)}\hat{N}\right\|_2^2 d\sigma} \end{aligned} . \tag{17}$$

Visibly, there indeed exists such a positive number $\varepsilon \in (0, \infty)$. Of course, the interval can be extended to $\left(-\left\|\vec{x}\hat{N}\right\|_2^2 \Big/ \int_{t_0}^t \left\|\vec{x}e^{(A\sigma)}\hat{N}\right\|_2^2 d\sigma, \infty\right)$, concretely, it is subject to the eigenvalues of matrix A.

C Proof of Theorem B.1

At first, establish the connection to matrix A with time variable t and assume an integral form for matrix P,

$$P = \int_0^t e^{\left(A^T\sigma\right)}\tilde{N}e^{(A\sigma)}d\sigma. \tag{18}$$

Then, we have that

$$\begin{aligned} &PA + A^T P \\ &= \int_0^t e^{\left(A^T\sigma\right)}\tilde{N}e^{(A\sigma)}Ad\sigma + \int_0^t A^T e^{\left(A^T\sigma\right)}\tilde{N}e^{(A\sigma)}d\sigma \\ &= \int_0^t \frac{d}{dt}\left(e^{\left(A^T\sigma\right)}\tilde{N}e^{(A\sigma)}\right)d\sigma \\ &= e^{\left(A^T\sigma\right)}\tilde{N}e^{(A\sigma)}\Big|_0^t \end{aligned} . \tag{19}$$

Let λ_i (i=1, $\cdots$, n) be the eigenvalues of matrix A with associated eigenvector v_i, thus A can be diagonalized as: $\Lambda = \hat{M}^{-1}A\hat{M}$, where $\hat{M} = [v_1, v_2, \cdots, v_n]$, $\Lambda = diag\{\lambda_1, \cdots, \lambda_n\}$.

Next, calculate $e^{(At)}$ using Taylor series,

$$\begin{aligned} e^{(At)} &= I + tA + \tfrac{t^2}{2!}A^2 + \cdots = \sum_{k=0}^{\infty}\tfrac{1}{k!}t^k A^k \\ &= MM^{-1} + tM\Lambda M^{-1} + \tfrac{t^2}{2!}M\Lambda^2 M^{-1} + \cdots \end{aligned} , \tag{20}$$

thus,

$$\begin{aligned} M^{-1}e^{(At)}M &= I + t\Lambda + \tfrac{t^2}{2!}\Lambda^2 + \cdots \\ &= \begin{pmatrix} e^{\lambda_1 t} & & 0 \\ & \ddots & \\ 0 & & e^{\lambda_n t} \end{pmatrix} = e^{(\Lambda t)} \end{aligned} . \tag{21}$$

Since all the eigenvalues of A have negative-real parts, thus we have that $e^{(At)} \to 0$ when $t \to \infty$, that is to say, through a long-time adversarial learning, $PA + A^T P$ will equal to $-\tilde{N}$. Due to the symmetry and positive-definite property of matrix $\tilde{N}$, it can be decomposed into the multiplication of one nonsingular matrix by its transpose[3], i.e. $\tilde{N} = \hat{N}^T\hat{N}$, which further indicates the symmetry of matrix P. Next, judge its definite property,

$$\begin{aligned} x^T Px &= \int_0^t x^T e^{\left(A^T\sigma\right)}\hat{N}^T\hat{N}e^{(A\sigma)}xdt \\ &= \int_0^t \left\|x^T e^{\left(A^T\sigma\right)}\hat{N}^T\right\|_2^2 d\sigma > 0 \end{aligned} . \tag{22}$$

Thus, there exists a unique positive definite solution P.

[3] [Chen (1999), Theorem 3.7]

D PROOF OF THEOREM B.2

We first introduce a theorem[4], namely, for a $p \times p$ rational positive-real and Hurwitz[5] matrix $Z(s)$, there exists an $r \times p$ Hurwitz matrix $Q(s)$ such that

$$\tilde{Z}(s) = Z(s) + Z^T(-s) = Q^T(-s)Q(s), \tag{23}$$

where $\hat{r}$ is the rank of $Z(s) + Z^T(-s)$ over the field of rational functions of s, and rank $Q(s)$=$\hat{r}$ for $Res[s] > 0$.

To prove the Theorem B.2, we need extend $\vec{y} = C\vec{x} + D\vec{u}$, where $\vec{u}$ can be seen as a flexible adjustment to the output $\vec{y}$ by design, it can be degraded to the original if D=0.

Let $\left\{\bar{\mathrm{A}}, \bar{\mathrm{B}}, \bar{\mathrm{C}}, \bar{\mathrm{D}}\right\}$ be a minimal realization[6] of positive-real function matrix $Z(s)$. Pursuant to Equation 23, there exists an $\hat{r} \times p$ function matrix $Q(s)$ satisfying it, and let $\{F, G, L, J\}$ be a minimal realization of $Q(s)$, thus, the minimal realization of $Q^T(s)$ is $\left\{-F^T, L^T, -G^T, J^T\right\}$. Then, the realization of $Q^T(-s) \cdot Q(s)$ can be furtherly inferred as:

$$\left\{\hat{\mathrm{A}}, \hat{\mathrm{B}}, \hat{\mathrm{C}}, \hat{\mathrm{D}}\right\} = \left\{\begin{pmatrix} F & 0 \\ L^T L & -F^T \end{pmatrix}, \begin{pmatrix} G \\ L^T J \end{pmatrix}, \left(J^T L, -G^T\right), J^T J\right\}. \tag{24}$$

Since both $Z(s)$ and $Z^T(-s)$ are Hurwitz, they have no common poles. Referring to the claim [(Duffin & Hazony, 1963), Theorem 7], we have $\delta\left[Z(s) + Z^T(-s)\right] = \delta[Z(s)] + \delta[Z^T(s)] = 2\delta[Z(s)]$, where $\delta[Z(\cdot)]$ denotes the degree of matrix $Z(\cdot)$. Therefore, $\left\{\hat{\mathrm{A}}, \hat{\mathrm{B}}, \hat{\mathrm{C}}, \hat{\mathrm{D}}\right\}$ is the minimal realization of $Q^T(-s)Q(s)$. From the claim [(Chen, 1999), Theorem 7.2], we know $\left(\hat{\mathrm{A}}, \hat{\mathrm{B}}\right)$ is controllable and $\left(\hat{\mathrm{A}}, \hat{\mathrm{C}}\right)$ is observable.

Given (F, L) is observable and F is Hurwitz, referring to Theorem B.1, there must exist a unique symmetric solution K, such that $KF + F^T K = -L^T L$. According to the proposition [(Ho & Kalman, 1966), Proposition 3], there must exist a nonsingular T to achieve a similar transformation from a minimal realization $\left\{\hat{\mathrm{A}}, \hat{\mathrm{B}}, \hat{\mathrm{C}}\hat{\mathrm{D}}\right\}$ to another minimal realization $\left\{\mathrm{T}\hat{\mathrm{A}}\mathrm{T}^{-1}, \mathrm{T}\hat{\mathrm{B}}, \hat{\mathrm{C}}\mathrm{T}^{-1}, \hat{\mathrm{D}}\right\}$. Use $T = \begin{pmatrix} I & 0 \\ K & I \end{pmatrix}$ to make a new minimal realization,

$$\begin{aligned} &\left\{\hat{\mathrm{A}}_1, \hat{\mathrm{B}}_1, \hat{\mathrm{C}}_1, \hat{\mathrm{D}}_1\right\} = \\ &\left\{\begin{pmatrix} F & 0 \\ 0 & -F^T \end{pmatrix}, \begin{pmatrix} G \\ KG + L^T J \end{pmatrix}, \left(J^T L + G^T K, -G^T\right), J^T J\right\} \end{aligned}. \tag{25}$$

Equivalently, we directly calculate the minimal realization of $Z(s) + Z^T(-s)$,

$$\begin{aligned} &\left\{\hat{\mathrm{A}}_2, \hat{\mathrm{B}}_2, \hat{\mathrm{C}}_2, \hat{\mathrm{D}}_2\right\} = \\ &\left\{\begin{pmatrix} \bar{A} & 0 \\ 0 & -\bar{A}^T \end{pmatrix}, \begin{pmatrix} \bar{B} \\ \bar{C}^T \end{pmatrix}, \left(\bar{C}, -\bar{B}^T\right), \bar{D} + \bar{D}^T\right\}. \end{aligned} \tag{26}$$

Given the eigenvalues of matrix A are all in the left-half plane (Hurwitz), and the eigenvalues of $-A^T$ is all in the right-half plane, indicating no common pole exists. Hence, we conclude the realization $\left\{\hat{\mathrm{A}}_2, \hat{\mathrm{B}}_2, \hat{\mathrm{C}}_2, \hat{\mathrm{D}}_2\right\}$ is minimal as well. Therefore, there must exist a nonsingular matrix R such that $\left\{\hat{\mathrm{A}}_1 = \mathrm{R}\hat{\mathrm{A}}_2\mathrm{R}^{-1}, \hat{\mathrm{B}}_1 = \mathrm{R}\hat{\mathrm{B}}_2, \hat{\mathrm{C}}_1 = \hat{\mathrm{C}}_2\mathrm{R}^{-1}, \hat{\mathrm{D}}_1 = \hat{\mathrm{D}}_2\right\}$. Suppose R is a block matrix as follows:

$$\begin{pmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{pmatrix}$$

[4] [Youla (1961), Theorem 2]

[5] All the eigenvalues are strictly in the left half plane.

[6] For a given $Z(s)$, there exist many sets $\left\{\bar{\mathrm{A}}, \bar{\mathrm{B}}, \bar{\mathrm{C}}, \bar{\mathrm{D}}\right\}$ to realize, however, there must be a minimal-dimension of matrix $\bar{\mathrm{A}}$, such a realization associated with minimal dimension $\bar{\mathrm{A}}$ is termed a minimal realization.

Thus, we can deduce that $-\mathrm{R}_{12}\bar{\mathrm{A}}^{\mathrm{T}}=\mathrm{FR}_{12}$. Then premultiplying by $\exp(Ft)$ and postmultiplying by $\exp\left(A^T t\right)$, then we have

$$\begin{array}{l}\frac{d}{dt}\left(\exp(Ft)\, R_{12}\exp(\bar{A}^T t)\right)= \\ \exp(Ft)\left[FR_{12}+R_{12}\bar{A}^T\right]\exp(\bar{A}^T t)=0\end{array}. \tag{27}$$

Thus, $\exp(Ft)\, R_{12}\exp(\bar{A}^T t)$ is a constant for all t. Furthermore, $R_{12}=\exp(Ft)\, R_{12}\exp(\bar{A}^T t)$ at t=0, also $R_{12}=\exp(Ft)\, R_{12}\exp(\bar{A}^T t)\to 0$ at $t\to\infty$, thus R_{12}=0. The sequel equations can be inferred, i.e., $F=R_{11}\bar{A}R_{11}^{-1}$, $G=R_{11}\bar{B}$, $J^T L+G^T K=\bar{C}R_{11}^{-1}$, $R_{22}^T G=\bar{B}$.

Assume $P=R_{11}^T K R_{11}$, $\hat{N}=LR_{11}$, $W=J$, then we can infer that

$$\begin{array}{c}P\bar{B}=\bar{C}^T-\hat{N}^T W \\ W^T W=\bar{D}+\bar{D}^T\end{array}. \tag{28}$$

On the basis of Eq. (4), we redefine a new transfer function $\breve{G}(s)$ as,

$$\begin{array}{l}\breve{G}(s)=M+(I+s\psi)\, G(s) \\ =M+C\left(sI-A\right)^{-1}B+\psi C\left(sI-A+A\right)\left(sI-A\right)^{-1}B \\ =M+\left(C+\psi CA\right)\left(sI-A\right)^{-1}B+\psi CB\end{array}. \tag{29}$$

Let λ_k be an eigenvalue of matrix A and v_k be the associated eigenvector. Then,

$$\left(C+\psi CA\right)v_k=\left(C+\psi C\lambda_k\right)v_k=\left(I+\lambda_k\psi\right)Cv_k. \tag{30}$$

The condition implies $(A,\, C)$ is observable. Naturally, Eq. (29) can be recognized as a quadri-tuple $\{A', B', C', D'\}$, namely $A'=A$, $B'=B$, $C'=(C+\psi CA)$, $D'=M+\psi CB$. If $\breve{G}(s)$ is strictly positive real and Hurwitz, referring to the conclusive Eq. (28), we can easily obtain the results $PB=-\hat{N}^T W+A^T C^T\psi+C^T$, $W^T W=\psi CB+2M+B^T C^T\psi$. That is to say, there exist a matrix W and a diagonal matrix ψ that make Eq. (15) correct.

E Deduction of one-dimensional mapping function

Let $\Gamma^{in}=\left\{\eta_i^{in}\middle|\,\forall i\in[1,N],\eta_i^{in}=\sum_{j=1}^{N}\tilde{B}_{(ij)}\right\}$ and $\vartheta_j(x_i)=\varphi(x_i,x_j)$, where $\tilde{B}$ is the adjacency matrix, then

$$\sum\nolimits_{j=1}^{N}\tilde{B}_{(ij)}\varphi(x_i,x_j)=\eta_i^{in}\left\langle\vartheta_j(x_i)\right\rangle, \tag{31}$$

where $\langle\vartheta_j(x_i)\rangle$ denotes the average of PNIB-caused perturbation by node i's neighbor j. Pursuant to Heterogeneous Mean-Field (HMF) approximation theory (Kundu et al., 2022), node i's in-neighbor j provides weight proportional to node j's out-degree, then we have the weighted average of $\vartheta_j(x_i)$ for node i

$$\left\langle\vartheta_j(x_i)\right\rangle_{HMF}=\frac{\frac{1}{N}\sum_{j=1}^{N}\eta_j^{out}\vartheta_j(x_i)}{\frac{1}{N}\sum_{j=1}^{N}\eta_j^{out}}, \tag{32}$$

where $\Gamma^{out}=\left\{\eta_j^{out}\middle|\,\forall j\in[1,N],\eta_j^{out}=\sum_{i=1}^{N}\tilde{B}_{(ij)}\right\}$ is the weighted out-degree of node j. As reported in Kundu et al. (2022), we can derive all the nodes have statistically identical perturbation affection if the degree correlation is small. Thus, the vector of adversarial perturbation can be approximated as

$$\Psi(\vartheta)=\frac{\mathbf{1}^T\tilde{B}\vartheta}{\mathbf{1}^T\tilde{B}1}=\frac{\frac{1}{N}\sum_{j=1}^{N}\eta_j^{out}\vartheta_j}{\frac{1}{N}\sum_{j=1}^{N}\eta_j^{out}}=\frac{\langle\Gamma^{out}\vartheta\rangle}{\langle\Gamma^{out}\rangle}. \tag{33}$$

From the angle of graph structure, we can presume the variance of components of node set (vectorized as $\vec{x}$) is small. Using HMF approximation theory, we rewrite Eq. (5) as

$$\frac{d\vec{x}}{dt}=\tilde{A}_{(\cdot)}\chi\left(\vec{x}\right)+\overrightarrow{\Gamma^{in}}\odot\varphi\left(\vec{x},\Psi\left(\vec{x}\right)\right), \tag{34}$$

where the sign $\odot$ denotes Hadamard product, i.e. multiply two vectors term by term. There are (N+1) variables with the join of $\Psi\left(\vec{x}\right)$,

$$\begin{aligned}\frac{d\Psi\left(\vec{x}\right)}{dt} &= \Psi\left(\tilde{A}_{(\cdot)}\chi\left(\vec{x}\right)+\vec{\Gamma}^{in}\odot\varphi\left(\vec{x},\Psi\left(\vec{x}\right)\right)\right)\\ &\approx \tilde{A}_{(\cdot)}\chi\left(\Psi\left(\vec{x}\right)\right)+\Psi\left(\vec{\Gamma}^{in}\right)\varphi\left(\Psi\left(\vec{x}\right),\Psi\left(\vec{x}\right)\right)\end{aligned}. \tag{35}$$

Thus, let $\tilde{x}=\Psi\left(\vec{x}\right)$, $\tilde{\beta}=\Psi\left(\vec{\Gamma}^{in}\right)$, we can infer the one-dimensional condensed Eq. (6).

F Proof of Theorem 3.2

Deriving from Theorem 2.5, we know that $\phi\left(\tilde{x}\right)=\frac{H\tilde{x}^2}{\theta\tilde{x}^2+1}$, and $\tilde{x}=\Psi\left(\vec{x}\right)$ is positive, thus $\phi\left(\tilde{x}\right)^T>0$, $\tilde{x}-M\phi\left(\tilde{x}\right)\geq 0$. Furtherly, $\frac{\phi(\tilde{x})}{\tilde{x}}=\frac{H\tilde{x}}{\theta\tilde{x}^2+1}\leq k$, therefore, let the first-order derivative equal to zero and compute the maximum value, we have $\frac{H\tilde{x}}{\theta\tilde{x}^2+1}\leq H\cdot\left(2\sqrt{\theta}\right)^{-1}$, thus $k\geq H\cdot\left(2\sqrt{\theta}\right)^{-1}$.

G Proof of Theorem 3.3

According to Theorem 2.5, we have $\breve{G}\left(s\right)=\frac{1}{k}+\left(1+s\gamma\right)G\left(s\right)=\frac{1}{k}+\left(1-\gamma a\right)\left(s+a\right)^{-1}\tilde{\beta}+\gamma\tilde{\beta}$. First, the $\det\left[\breve{G}(s)+\breve{G}^T\left(-s\right)\right]$ is obviously not identically zero. Since a is positive, then $\breve{G}(s)$ is Hurwitz, that is, the poles of all elements of $\breve{G}(s)$ have negative real parts. Pursuant to the claim [Khalil (2002), Theorem 6.1], we calculate

$$\breve{G}\left(j\omega\right)+\breve{G}^T\left(-j\omega\right)=\frac{2}{k}+\frac{2\tilde{\beta}\left(a+\gamma\omega^2\right)}{a^2+\omega^2}. \tag{36}$$

Thereby, it is positive definite for all $\omega\in\mathbb{R}$, and $\breve{G}\left(s\right)$ is strictly positive real, which further implies Equation 8 has asymptotically-stable equilibrium point.

H Theoretical Perturbed-Bound

In theory, we can quantify the perturbation bound through the following two steps:

Step I: Theorem 3.2 in fact gives the perturbation bounds, wherein the constant k, required to be in the interval $\left[H\cdot\left(2\sqrt{\theta}\right)^{-1},\infty\right)$, reflects the perturbation bounds (Theorem 2.5) when ASEP exists. As we know, for each dataset the parameter H is the average in-degree weight, and directly subject to the perturbation budgets ρ, thus we can assume the mapping function as: $H=P\left(\rho\right)$. Given the concrete dataset, we have the value θ according to Table 4. Therefore, we can further infer the detailed interval $\left[P\left(\rho\right)\cdot\left(2\sqrt{\theta}\right)^{-1},\infty\right)$ that constrains the perturbation bounds.

Step II: According to Theorem 2.5, for one-dimensional function, the expression $\frac{1}{k}+\left(1+s\gamma\right)G\left(s\right)$ is required to be strictly positive, and assume $G\left(s\right)=G(\varpi j)$, we can infer $\frac{1}{k}+\mathrm{Re}\left[G\left(\varpi j\right)\right]-\gamma\varpi\mathrm{Im}\left[G\left(\varpi j\right)\right]>0$, $\varpi\in[0,\infty]$, where $G(\varpi j)=\mathrm{Re}\left[G\left(\varpi j\right)\right]+j\mathrm{Im}\left[G\left(\varpi j\right)\right]$. If we regard $\mathrm{Re}\left[G\left(\varpi j\right)\right]$ and $\varpi\mathrm{Im}\left[G\left(\varpi j\right)\right]$ as two coordinates x and y with ϖ as a parameter, the intermediate variation function (a.k.a. bounded perturbation) must be below the line $\frac{1}{k}+x-\gamma y=0$, intercepting the point $-\frac{1}{k}+0j$ in complex plane.

Therefore, from the two steps above, we can quantify the bounds in relation to perturbation budget.

I IMPLEMENTATION ALGORITHMS

Algorithm 1: Graph Parameters Calculation

Input: Adjacency matrix $\widehat{A}$
Output: Graph state parameters $\tilde{\beta}$, $\tilde{x}$

Procedure calculate_parameters ($\widehat{A}$)
inDeg ← sum($\widehat{A}$, axis=0)
outDeg ← sum($\widehat{A}$, axis=1)
degCentrality ← (inDeg + outDeg) / (len($\widehat{A}$)-1)
$\tilde{\beta}$ ← sum(outDeg × inDeg) / sum(outDeg)
$\tilde{x}$ ← sum(outDeg × degCentrality) /sum(outDeg)
return $\tilde{\beta}$, $\tilde{x}$

Algorithm 2: Adjacency Matrix Modification

Input: Adjacency matrix $\widehat{A}$, rate of adversarial perturbation $\widehat{RoP}$
Output: Resilience_matrix $\widehat{RA}$

Procedure defense ()
simi_list ← calculate_edge_simi($\widehat{A}$)
$\widehat{RA}$ ← $\widehat{A}$
critical_$\tilde{\beta}$ ← 0
critical_$\tilde{x}$ ← 0
Initial_lenth ← len(simi_list)
while *True* **do**
    modified_matrix, simi_list ← remove_low_simi_edges($\widehat{RA}$.copy(), simi_list)
    $\tilde{\beta}$, $\tilde{x}$ ← calculate_parameters(modified_matrix)
    if $\tilde{\beta} >$ *critical_*$\tilde{\beta}$ *and* $\tilde{x} >$ *critical_*$\tilde{x}$ **then**
        critical_$\tilde{\beta}$ ← $\tilde{\beta}$
        critical_$\tilde{x}$ ← $\tilde{x}$
        $\widehat{RA}$ ← modified_matrix
    if *len(simi_list)* $\leq$ *Initial_lenth* × $\widehat{RoP}$ **then**
        break
return $\widehat{RA}$

The inferred asymptotically-stable equilibrium point (ASEP) servers as the referral to purifying the perturbed graph by enforcing adjacency matrix-projection coordinator to be close to such equilibrium point. The implementation mainly involves two phases: i) graph parameters calculation (Algorithm 1). This part aims to compute the two important parameters $\tilde{\beta}$ and $\tilde{x}$ as the referral to modify the adjacency matrix of perturbed graph; and ii) adjacency matrix modification (Algorithm 2). Referring to the two calculated critical-state parameters, this part mainly achieves the modification on adversarial edges. Specifically, we first define function *calculate_edge_simi()* to empirically search the neighboring nodes within h-hop (h=2), then use Jaccard metric to calibrate the pairwise similarity for each two nodes. Based on the similarity, we define function *remove_low_simi_edges()* to remove those edges with low similarity during the "while" loop, until the inferred coordinator point from the iteratively-modified adjacency matrix approaches to the "Original" point (i.e. ASEP).

J EXPERIMENT & PERFORMANCE

J.1 CONFIGURATION

Datasets. Our experiments are performed on five commonly-used scalable realistic graph datasets: Polblogs, Cora, Cora_ML, Citeseer, and Amazon Photo. The statistics are sketched in Table 3. Given our method aims to promote the adversarial resilience from the perspective of graph topology, thus, the dataset Polblogs is completely adaptive since the node feature is unavailable. To sufficiently verify the effectiveness of our proposed method, we also introduce another four feature-involved datasets, however, our method and baselines only concentrate the graph topology information but disregard the node feature in the experiments. Concretely, we set the discrepancy of pairwise-node features as the same in our experimental procedure.

Table 3: Dataset statistics

Dataset	# Node	# Edge	# Feature	# Classi.
PolBlogs	1222	16714	/	2
Cora_ML	2810	7981	2879	7
Cora	2485	5069	1433	7
Citeseer	2110	3668	3703	6
Amazon Photo	7650	238162	745	8

Baselines. Methodologically, our work aims to improve the graph's adversarial resilience through modifying its adjacency matrix, same as ours, GCN-SVD and GCN also focus on the adjacency-matrix optimization as well. On the other hand, for other non-adjacency-matrix based defense approaches, such as GAT, RGCN, HANG, Mid-GCN, and GOOD-AT, our approach can also boost them from the lens of graph topology, that is, our work can significantly enhance the adversarial resilience by constructing a proper graph structure in advance.

- **GCN (Kipf & Welling, 2017):** Graph Convolutional Network (GCN), as a basic and prevalent deep learning model, is good at handling network/graph-specified data.
- **GAT (Velickovic et al., 2018):** Graph Attention Network (GAT) equipped with attention layers aims to learn differential weights to different neighbors for individuals. It is usually utilized to defend against various graph adversarial attacks.
- **RGCN (Zhu et al., 2019):** It first employs Gaussian distributions as hidden representations of nodes to absorb adversarial changes in the variances of the Gaussian distribution, then assigns different weights to node neighborhoods pursuant to their variances during the convolutions.
- **GCN-SVD (Entezari et al., 2020):** It is a preprocessing method to withstand adversarial attacks by decomposing adjacency matrix of the perturbed graph and then yielding a low-rank approximate matrix to purify the adversarial perturbations (edges).
- **HANG (Zhao et al., 2023):** It is a recently-proposed graph neural flow method, and resorting to the ordinary differential equation to learn a robust representation embeddings for graph nodes.
- **Mid-GCN (Huang et al., 2023):** It is a novel method recently proposed to leverage the mid-frequency signals (Laplacian eigenvalue around 1) on graphs for the promotion of robustness through designing a mid-pass filtering GCN model.
- **GOOD-AT (Li et al., 2024):** Upon the incorporation of OOD detection, the authors leverage popular adversarial training paradigm to defend against poisoning and evasion attacks.

Graph Adversarial Attacks. As we know, the distinction between evasion attack and poisoning attack lies in different phases, i.e. the former works in inference process by modifying the testing samples, while the latter works in training phase by modifying the training samples. Thus, the following non-targeted and targeted GAAs used in this paper belong to poisoning attacks:

- **Metattack (Zügner & Günnemann, 2019):** It utilizes meta-learning to generate adversarial perturbations by computing gradients to determine how to modify the graph structure

in order to achieve the attack objective. The gradient computation is typically based on the loss function of the target nodes, aiming to identify which edges should be modified to maximally affect the GNN's prediction results. Based on the computed gradient information, the attacker makes small changes to the graph, such as adding new edges or deleting existing ones. The modified graph structure is then used to retrain the GNN model and evaluate the success of the attack.

- **CE-PGD (Xu et al., 2019):** It uses the negative cross-entropy loss function to measure the difference between the model's predictions and the true labels. The attack starts from the original graph by randomly adding or deleting edges to initialize the adversarial graph. In each iteration, it computes the gradient of the loss function with respect to the current adversarial graph, then updates the graph using the gradient descent method. The updated graph structure is then projected back into the feasible region defined by the attack constraints.

- **DICE (Waniek et al., 2016):** It is a heuristic adversarial attack inspired by modularity, which is used to measure the quality of any given community structure. In other words, it promotes structures where communities have dense connections within themselves and sparse connections between different communities. The algorithm mainly consists of two steps. The first step reduces the connection density among nodes within the same region. The second step increases the connections between nodes in that region and nodes in other regions, by which the algorithm can ignore that region, meaning it will not identify it as a separate community, but instead assign the nodes within that region to other communities.

- **Adaptive Attack:** It is specifically-designed for our approach, i.e. relax the adjacency to the interval $[0, 1]^{n\times n}$ instead of the original element (0 or 1), which would severely cause incorrect calculation for the two key variables $\tilde{\beta}$ and $\tilde{x}$ in our method, further leading to inaccurate location of asymptotically-stable equilibrium point.

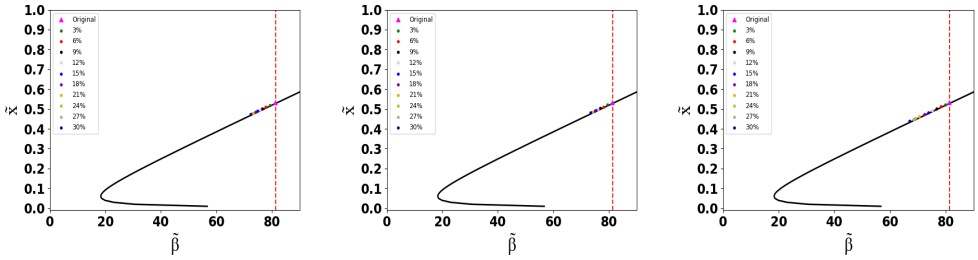
Execution Settings. For each dataset, 10% nodes are randomly chosen for training, 10% nodes for verifying and the rest 80% nodes for testing. For GCN, GAT, HANG, and Mid-GCN, the parameters are kept same as the default settings in their original papers. For GCN-SVD, the reduced rank is tuned from {20, 40, 60, 80, 100}. For our EquiliRes, the ratio of edge modification (deletion/addition) over dissimilar noes is tuned from {1%, 5%, 10%, 15%}, here 1% just for the clean graph, i.e. adversarial perturbation rate is 0%. The experimental results are calculated on average for 10 times. Additionally, the running environment is: ubuntu 20.04.2, GPU: GeForce RTX 3090.

J.2 Equilibrium-Point Trajectory

Pursuant to the parameters as listed in Table 4, we draw the trajectories of asymptotically-stable equilibrium-points on datasets Polblogs, Amazon Photo, Cora_ML, Cora, and Citeseer as displayed in Fig. 2–6, wherein the horizontal-longitudinal coordinate converts. From the trajectories, besides the analogous analytics and reviews as presented in Section 5.2, we also obtain the following observations: i) although the plot of adversarial resilience can well-match the attack effect under different rates of adversarial perturbation, the graph's adversarial resilience indeed declines as the strength of perturbation enlarges, which makes sense in intuition, you can image an extreme case that if all edges are removed, the nodes would become isolated and cannot be classified without the connections to each other; and ii) different datasets have distinct trajectories of adversarial resilience, stemming from the difference of graph's internal topology. Thereby, we need carefully figure out such an equilibrium-point trajectory for each unique graph.

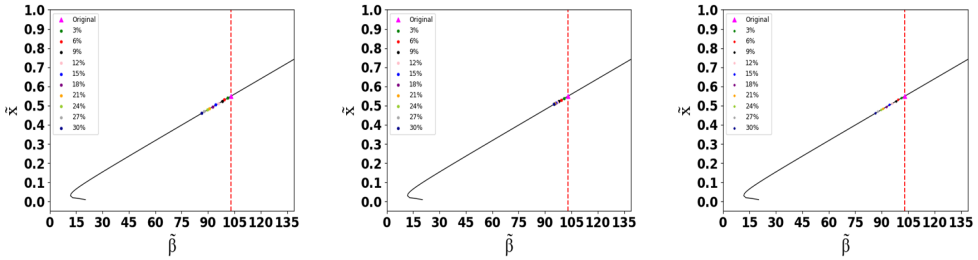
J.3 Performance on Attack Resilience

Table 5 and Table 6 show the comparative results separately between our approach and another two adjacency-matrix optimization defense methods GCN and GCN-SVD. Table 7 exhibits the results of GAT, RGCN, HANG, Mid-GCN, and GOOD-AT from the angle of combination with our EquiliRes, from which we can gain the analogous analysis and conclusion as stated in Section 5.3.1.



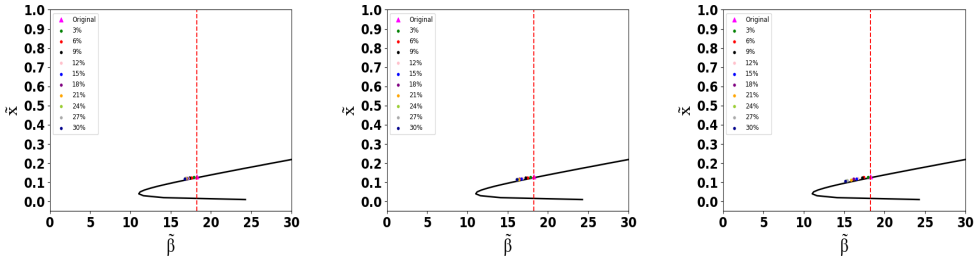
(a) Metattack (b) CE-PGD (c) DICE

Figure 2: Adversarial effect with Polblogs.



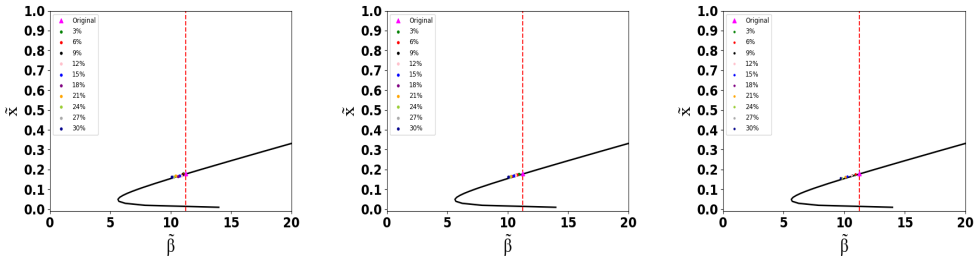
(a) Metattack (b) CE-PGD (c) DICE

Figure 3: Adversarial effect with Amazon Photo.



(a) Metattack (b) CE-PGD (c) DICE

Figure 4: Adversarial effect with Cora_ML.

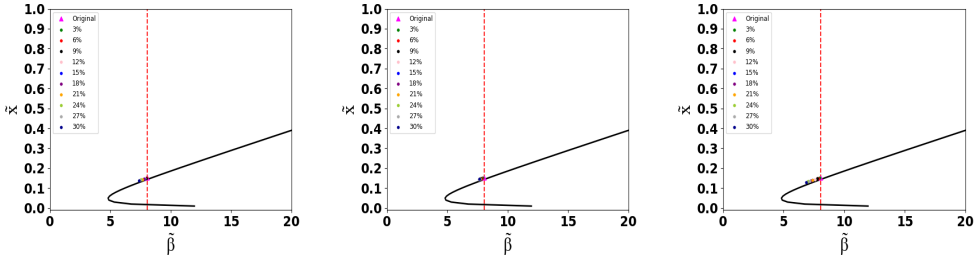


(a) Metattack (b) CE-PGD (c) DICE

Figure 5: Adversarial effect with Cora.

Table 4: Parameters for generating equilibrium-point trajectory

Dataset	GAA	Parameter
Polblogs	Metattack/CE-PGD/DICE	$\theta = 16.6^2$ η=0.45
Cora_ML	Metattack/CE-PGD/DICE	$\theta = 24^2$ η=10
Cora	Metattack/CE-PGD/DICE	$\theta = 21^2$ η=10
Citeseer	Metattack/CE-PGD/DICE	$\theta = 21^2$ η=10
Amazon Photo	Metattack/CE-PGD/DICE	$\theta = 31.8^2$ η=8



(a) Metattack (b) CE-PGD (c) DICE

Figure 6: Adversarial effect with Citeseer.

J.4 Adaptive Attack

To comprehensively evaluate the performance, we also introduce a specially-designed Adaptive Attack for our approach, i.e. relax the adjacency to the interval $[0, 1]^{n\times n}$ instead of the original binary element (0 or 1), which severely causes incorrect calculation for the two key variables $\tilde{\beta}$ and $\tilde{x}$ in our method, further leading to inaccurate location of asymptotically-stable equilibrium point. From the experimental results reported in Table 8, we can observe that although the performance under Adaptive Attack is remarkably degraded compared to that under other adversarial attacks like Metattack in Table 2, EquiliRes still can retain a comparative performance. This also reflects that the designed decimal-based Adaptive Attack is inapplicable for GAT and Mid-GCN, but applicable for our EquiliRes and HANG.

To sum up, we can highlight three aspects from the experiments: i) our proposed method can significantly promote the adversarial resilience from the single perspective of graph topology; ii) our work

Table 5: Accuracy (%) of node classification on Citeseer

GAA	Def. \ Per.	0%	5%	10%	15%	20%	25%
Metattack	GCN	54.85±1.71	52.30±2.13	46.92±1.58	43.30±1.42	41.18±1.27	39.40±0.73
	GCN-SVD	46.12±1.54	45.30±1.64	44.89±1.44	44.28±1.84	43.26±1.45	42.84±1.48
	EquiliRes	**55.76±2.07** (**0.91%**↑)	**55.05±1.21** (**2.75%**↑)	**52.59±1.04** (**5.67%**↑)	**50.79±1.78** (**6.51%**↑)	**48.19±0.34** (**4.93%**↑)	**45.15±2.19** (**2.31%**↑)
CE-PGD	GCN	54.85±1.71	52.76±1.95	50.46±2.44	50.34±1.79	49.25±1.51	49.68±1.46
	GCN-SVD	46.12±1.54	46.22±1.78	46.16±1.69	45.73±1.93	45.80±3.20	47.70±3.15
	EquiliRes	**55.76±2.07** (**0.91%**↑)	**54.67±2.44** (**1.91%**↑)	**53.11±2.41** (**2.65%**↑)	**52.87±2.35** (**2.53%**↑)	**51.55±2.32** (**2.30%**↑)	**52.89±2.18** (**3.21%**↑)
DICE	GCN	54.85±1.71	51.48±1.48	51.16±1.75	49.00±1.43	47.71±1.48	45.94±1.53
	GCN-SVD	46.12±1.54	44.70±1.96	41.65±3.07	41.58±2.43	39.61±2.84	37.71±1.75
	EquiliRes	**55.76±2.07** (**0.91%**↑)	**53.30±2.09** (**1.82%**↑)	**51.42±1.73** (**0.26%**↑)	**50.81±2.16** (**1.81%**↑)	**49.46±1.78** (**1.75%**↑)	**47.01±1.68** (**1.07%**↑)

Table 6: Accuracy (%) of node classification on Amazon Photo

GAA	Def. \ Per.	0%	5%	10%	15%	20%	25%
Metattack	GCN	92.15±0.34	82.42±7.74	82.81±5.81	79.85±13.16	62.95±14.96	61.08±15.26
	GCN-SVD	87.13±0.47	84.32±7.94	82.01±9.81	72.39±12.46	71.21±12.29	63.57±16.78
	EquiliRes	**92.87±0.46** (**0.72%**↑)	**91.65±0.21** (**7.33%**↑)	**85.11±20.05** (**2.30%**↑)	**82.25±30.00** (**2.40%**↑)	**74.35±31.92** (**3.14%**↑)	**81.14±25.58** (**17.57%**↑)
CE-PGD	GCN	92.15±0.34	90.42±0.37	90.15±0.31	90.12±0.29	88.41±0.56	88.29±0.34
	GCN-SVD	87.13±0.47	85.82±4.83	87.75±0.36	86.78±2.47	86.75±2.17	86.17±0.49
	EquiliRes	**92.87±0.46** (**0.72%**↑)	**91.12±0.32** (**0.70%**↑)	**90.62±0.45** (**0.47%**↑)	**90.33±0.17** (**0.21%**↑)	**89.39±0.46** (**0.98%**↑)	**88.58±0.25** (**0.29%**↑)
DICE	GCN	92.15±0.34	90.96±0.65	89.03±5.30	89.00±1.01	86.60±3.17	84.90±3.18
	GCN-SVD	87.13±0.47	85.03±0.56	82.74±0.89	81.54±0.41	75.75±2.72	76.14±0.57
	EquiliRes	**92.87±0.46** (**0.72%**↑)	**91.09±0.33** (**0.13%**↑)	**91.45±0.21** (**2.42%**↑)	**90.84±0.12** (**1.84%**↑)	**89.45±0.53** (**2.85%**↑)	**88.59±0.11** (**3.69%**↑)

also enables to benefit the existing defense methods to resist adversarial attacks through building an equilibrium state in advance for the protected graph; and iii) even if under Adaptive Attack, our approach still can perform comparatively with baselines.

J.5 Ablations

With/Without ASEP. We run experiments with/without referring to ASEP as guidance on Polblogs, Cora_ML, and Cora. From the experimental results in Tables 9–11, we can observe the performance with ASEP as a guidance significantly outperforms that without it.

More Attack Budget. Furthermore, we also run a set of experiments to demonstrate whether the derived equilibrium points still well-match the theoretical ASEP curve with larger attack budget (APR) from 30% to 45%, the results on Polblogsare under Metattack are shown in Table 12, from which we observe that the derived equilibrium points $(\tilde{\beta}, \tilde{x})$ can appropriately locate at the theoretical ASEP curve in Fig. 2(a).

J.6 Rank of Adjacency Matrix

The recent research points out the GAAs could enlarge the rank of adjacency matrix quickly, thus, to validate the effectiveness of our EquiliRes, we produce two groups of comparative experiments under the three adversarial attacks, as shown in Fig. 7 on Polblogs and Fig. 8 on Citeseer, to observe the variation of adjacency matrix rank before and after our defense method is engaged.

Seen from the experimental results, we have the following observations: i) as the APR increases, the rank value goes up gradually under the three adversarial attacks on dataset Polblogs, especially with visible lift for Citeseer dataset, this phenomenon shows that graph adversarial attack renders the rank growth to establish connection between dissimilar nodes; ii) correspondingly, our EquiliRes enables to dramatically decline the rank to remove the adversarial edges; and iii) the rank-declining ratio by our proposed EquiliRes has a descending tendency as the perturbation strength becomes larger, this matches the experimental results in Tables 1–2 and Tables 5–7 , reflecting the accuracy of node-classification task would become lower under the gradually growing adversarial perturbations.

J.7 Singular-Value Variation

Jin et al. (2020) disclose the attack Metattack can enlarge the singular values of adjacency matrix, thus we run a set of experiments to showcase the variation of singular values of adjacency matrix after using our EquiliRes on moderate-scale Polblogs and large-scale Amazon Photo datasets, and depict the results in Fig. 9 and Fig. 10, from which the following observations can be obtained.

First, our EquiliRes indeed reduces the singular values of adjacency matrices under the three adversarial attacks on both datasets, which indicates our proposed equilibrium-point based modification

Table 7: Accuracy (%) of node classification under combination with EquiliRes using Metattack

Dataset	Def. \ Per.	0%	5%	10%	15%	20%	25%
Cora_ML	GAT	77.55±1.15	68.86±1.04	64.13±1.37	60.65±0.92	58.75±1.70	56.43±1.26
	Ours+GAT	**77.69±1.30** (**0.14%**↑)	**71.15±1.85** (**2.29%**↑)	**68.32±1.99** (**4.19%**↑)	**66.11±1.61** (**5.46%**↑)	**64.52±1.71** (**5.77%**↑)	**62.12±1.59** (**5.69%**↑)
	RGCN	80.57±0.18	70.43±0.32	62.38±0.29	55.78±0.34	52.14±0.24	48.10±0.28
	Ours+RGCN	**80.98±0.14** (**0.41%**↑)	**71.48±0.27** (**1.05%**↑)	**66.04±0.32** (**3.66%**↑)	**59.87±0.23** (**4.09%**↑)	**57.07±0.19** (**4.93%**↑)	**52.62±0.23** (**4.52%**↑)
	HANG	64.11±4.49	57.98±2.30	52.98±2.23	50.54±2.52	47.64±2.51	45.93±2.93
	Ours+HANG	**64.34±5.03** (**0.23%**↑)	**60.47±3.58** (**2.49%**↑)	**57.08±2.68** (**4.10%**↑)	**55.34±3.14** (**4.80%**↑)	**54.23±3.63** (**6.59%**↑)	**53.06±3.25** (**7.13%**↑)
	Mid-GCN	83.40±0.21	80.01±0.37	78.95±0.37	77.80±0.55	75.66±0.57	74.17±0.72
	Ours+Mid-GCN	**83.49±0.16** (**0.09%**↑)	**80.22±0.46** (**0.21%**↑)	**79.70±0.34** (**0.75%**↑)	**78.03±0.63** (**0.23%**↑)	**76.19±0.72** (**0.53%**↑)	**74.53±0.83** (**0.36%**↑)
Cora	GAT	73.28±1.43	67.53±1.39	61.20±1.77	58.09±1.21	55.07±1.59	52.16±1.75
	Ours+GAT	**73.43±1.36** (**0.15%**↑)	**67.65 ±1.60** (**0.12%**↑)	**63.53±2.06** (**2.33%**↑)	**60.08±2.29** (**1.99%**↑)	**57.66±2.22** (**2.59%**↑)	**55.23±2.05** (**3.07%**↑)
	RGCN	77.09±0.29	72.57±0.26	65.75±0.28	57.26±0.56	55.67±0.35	49.33±0.46
	Ours+RGCN	**77.45±0.37** (**0.36%**↑)	**73.35±0.18** (**0.78%**↑)	**65.87±0.47** (**0.12%**↑)	**58.57±0.49** (**1.31%**↑)	**56.03±0.43** (**0.36%**↑)	**49.95±0.38** (**0.62%**↑)
	HANG	58.22±4.79	55.23±3.80	50.25±3.30	48.36±3.33	45.84±3.12	44.54±3.14
	Ours+HANG	**58.40±4.82** (**0.18%**↑)	**55.49±3.98** (**0.26%**↑)	**52.41±3.50** (**2.16%**↑)	**50.22±3.58** (**1.86%**↑)	**48.77±2.99** (**2.93%**↑)	**47.29±2.85** (**2.75%**↑)
	Mid-GCN	83.10±0.41	75.81±15.82	78.88±0.93	75.43±1.11	75.50±1.35	74.43±0.83
	Ours+Mid-GCN	**83.23±0.36** (**0.13%**↑)	**81.27±0.72** (**5.46%**↑)	**80.20±1.10** (**1.32%**↑)	**77.69±0.82** (**2.26%**↑)	**76.74±0.64** (**1.24%**↑)	**75.16±0.76** (**0.73%**↑)
	GOOD-AT	84.12±0.34	82.73±0.78	81.31±1.65	80.20±1.24	79.15±0.36	78.88±0.85
	Ours+GOOD-AT	**84.30±0.56** (**0.18%**↑)	**83.21±0.63** (**0.48%**↑)	**81.99±0.65** (**0.68%**↑)	**81.32±0.41** (**1.12%**↑)	**80.30±0.43** (**1.15%**↑)	**80.46±0.34** (**1.58%**↑)
Citeseer	GAT	64.04±1.85	60.21±1.91	54.00±1.81	49.92±2.02	46.96±1.70	44.65±0.97
	Ours+GAT	**64.17±1.73** (**0.13%**↑)	**61.14±1.97** (**0.93%**↑)	**54.60±1.65** (**0.60%**↑)	**53.21±1.36** (**3.29%**↑)	**51.30±1.64** (**4.34%**↑)	**48.39±2.31** (**3.74%**↑)
	RGCN	67.53±0.49	63.49±0.35	57.22±0.39	51.91±0.33	50.37±0.65	45.34±0.70
	Ours+RGCN	**67.72±0.58** (**0.19%**↑)	**63.50±0.43** (**0.01%**↑)	**58.10±0.32** (**0.88%**↑)	**51.98±0.34** (**0.07%**↑)	**50.49±0.53** (**0.12%**↑)	**45.63±0.47** (**0.29%**↑)
	HANG	56.48±2.85	52.92±1.91	48.05±2.72	45.12±0.83	44.29±1.83	40.61±1.96
	Ours+HANG	**56.54±2.74** (**0.06%**↑)	**53.17±2.37** (**0.25%**↑)	**49.62±2.17** (**1.57%**↑)	**49.62±2.17** (**4.50%**↑)	**46.64±1.60** (**2.35%**↑)	**44.70±2.40** (**4.09%**↑)
	Mid-GCN	74.52±0.32	72.37±0.67	71.05±0.66	65.84±1.06	67.00±0.52	63.06±0.99
	Ours+Mid-GCN	**74.66±0.47** (**0.14%**↑)	**72.70±0.37** (**0.33%**↑)	**71.52±0.71** (**0.47%**↑)	**66.45±1.03** (**0.61%**↑)	**67.84±0.50** (**0.84%**↑)	**63.84±1.25** (**0.78%**↑)
	GOOD-AT	74.33±0.23	71.71±0.55	70.17±1.28	68.88±0.84	68.22±1.60	66.58±2.68
	Ours+GOOD-AT	**74.51±0.27** (**0.18%**↑)	**72.41±0.27** (**0.70%**↑)	**70.62±2.07** (**0.45%**↑)	**70.50±1.31** (**1.62%**↑)	**70.03±1.72** (**1.81%**↑)	**68.12±2.75** (**1.54%**↑)
Amazon Photo	GAT	93.54±0.18	91.73±0.39	76.53±31.14	69.99±32.95	51.57±38.85	62.53±33.59
	Ours+GAT	**93.70±0.16** (**0.16%**↑)	**91.74±0.18** (**0.01%**↑)	**83.65±20.01** (**7.12%**↑)	**76.19±29.97** (**6.20%**↑)	**74.30±31.89** (**22.73%**↑)	**81.05±25.55** (**18.52%**↑)
	HANG	93.47±0.32	91.62±0.54	91.51±0.39	90.93±0.68	90.16±0.56	89.92±0.61
	Ours+HANG	**93.63±0.43** (**0.16%**↑)	**92.64±0.44** (**1.02%**↑)	**91.72±0.35** (**0.21%**↑)	**91.97±0.42** (**1.04%**↑)	**91.04±0.39** (**0.88%**↑)	**90.85±0.73** (**0.93%**↑)
	Mid-GCN	77.25±0.48	74.39±0.49	65.95±0.82	57.08±1.70	48.75±0.88	46.75±1.13
	Ours+Mid-GCN	**78.45±0.50** (**1.20%**↑)	**74.81±0.82** (**0.42%**↑)	**66.15±1.14** (**0.20%**↑)	**59.47±1.18** (**2.39%**↑)	**51.02±1.35** (**2.27%**↑)	**47.37±1.43** (**0.62%**↑)

Table 8: Accuracy (%) of node classification under Adaptive Attack

Dataset	Def. \ Per.	5%	10%	15%	20%	25%
Polblogs	GAT	88.14±1.03	79.17±2.27	73.96±1.64	70.24±1.76	69.06±0.77
	Ours+GAT	83.89±0.70	75.49±1.41	72.07±2.09	70.58±1.56	69.01±0.96
	HANG	79.78±2.25	75.27±1.96	74.59±2.04	71.31±5.57	72.63±3.48
	Ours+HANG	79.62±2.12	75.64±1.43	73.12±3.79	72.55±2.89	71.06±3.57
	Mid-GCN	88.29±2.92	87.10±3.27	84.56±2.96	79.87±4.63	78.99±5.15
	Ours+Mid-GCN	87.82±2.82	86.72±2.72	83.07±4.57	80.24±4.82	80.24±4.82

Table 9: Accuracy (%) of node classification with/without referring to ASEP on Polblogs

GAA	Per.	0%	5%	10%	15%	20%	25%
Metattack	w/o	93.15±0.26	78.87±1.08	72.39±1.25	69.24±0.74	69.53±0.99	68.08±0.30
	w/	**94.34±0.81** (**1.19%**↑)	**90.11±2.62** (**11.24%**↑)	**89.70±2.22** (**17.31%**↑)	**86.53±1.47** (**17.29%**↑)	**81.41±1.94** (**11.88%**↑)	**73.11±2.60** (**5.03%**↑)
CE-PGD	w/o	93.15±0.26	84.14±0.15	79.01±0.14	77.02±0.11	75.43±0.08	73.74±0.07
	w/	**94.34±0.81** (**1.19%**↑)	**86.69±1.19** (**2.55%**↑)	**82.67±1.12** (**3.66%**↑)	**79.93±1.39** (**2.91%**↑)	**76.82±1.01** (**1.39%**↑)	**75.59±1.41** (**1.85%**↑)
DICE	w/o	93.15±0.26	86.74±0.42	85.63±0.72	81.61±0.79	80.15±0.38	76.50±0.52
	w/	**94.34±0.81** (**1.19%**↑)	**89.47±0.95** (**2.73%**↑)	**88.43±2.32** (**2.80%**↑)	**86.85±2.11** (**5.24%**↑)	**84.75±1.54** (**4.60%**↑)	**83.45±1.63** (**6.95%**↑)

Table 10: Accuracy (%) of node classification with/without referring to ASEP on Cora_ML

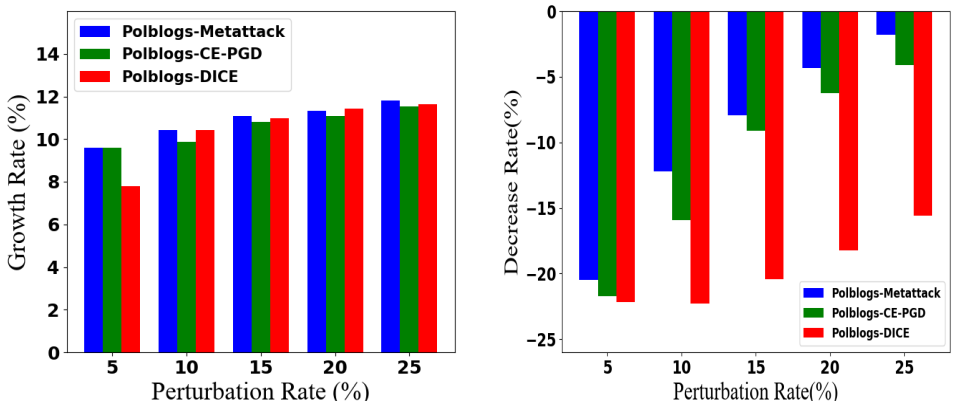
GAA	Per.	0%	5%	10%	15%	20%	25%
Metattack	w/o	74.31±0.50	67.15±1.00	59.42±0.66	53.83±0.97	49.98±0.66	46.49±0.92
	w/	**74.77±1.49** (**0.46%**↑)	**69.12±1.12** (**1.97%**↑)	**65.62±0.96** (**6.20%**↑)	**62.70±1.32** (**8.87%**↑)	**60.72±0.61** (**10.74%**↑)	**58.39±1.12** (**11.90%**↑)
CE-PGD	w/o	74.31±0.50	70.81±0.36	69.73±0.33	69.72±0.30	68.76±0.27	68.28±0.32
	w/	**74.77±1.49** (**0.46%**↑)	**72.77±1.37** (**1.96%**↑)	**72.65±1.26** (**2.92%**↑)	**72.65±2.27** (**2.93%**↑)	**71.98±2.80** (**3.22%**↑)	**71.70±2.30** (**3.42%**↑)
DICE	w/o	74.31±0.50	70.63±0.38	48.77±0.28	66.83±0.65	64.34±0.41	61.17±0.33
	w/	**74.77±1.49** (**0.46%**↑)	**72.43±1.83** (**1.80%**↑)	**69.53±1.38** (**20.76%**↑)	**67.52±1.38** (**0.69%**↑)	**65.74±1.30** (**1.40%**↑)	**63.58±1.58** (**2.41%**↑)

Table 11: Accuracy (%) of node classification with/without referring to ASEP on Cora

GAA	Per.	0%	5%	10%	15%	20%	25%
Metattack	w/o	62.05±0.83	61.30±0.56	56.44±1.26	51.44±0.54	51.04±0.39	46.01±0.55
	w/	**67.74±2.65** (**5.69%**↑)	**64.26±0.61** (**2.96%**↑)	**59.02±1.28** (**2.58%**↑)	**57.31±2.05** (**5.87%**↑)	**54.78±0.77** (**3.74%**↑)	**52.04±0.82** (**6.03%**↑)
CE-PGD	w/o	62.05±0.83	65.15±1.21	64.37±0.63	62.31±1.18	61.07±0.83	60.61±0.92
	w/	**67.74±2.65** (**5.69%**↑)	**67.55±3.69** (**2.40%**↑)	**65.64±3.81** (**1.27%**↑)	**63.29±3.11** (**0.98%**↑)	**62.81±3.24** (**1.74%**↑)	**62.27±3.69** (**1.66%**↑)
DICE	w/o	62.05±0.83	59.35±1.01	58.07±0.71	54.61±0.56	53.14±1.05	51.60±0.86
	w/	**67.74±2.65** (**5.69%**↑)	**66.67±2.80** (**7.32%**↑)	**63.58±3.13** (**5.51%**↑)	**61.77±2.95** (**7.16%**↑)	**60.41±3.17** (**7.27%**↑)	**57.78±2.63** (**6.18%**↑)

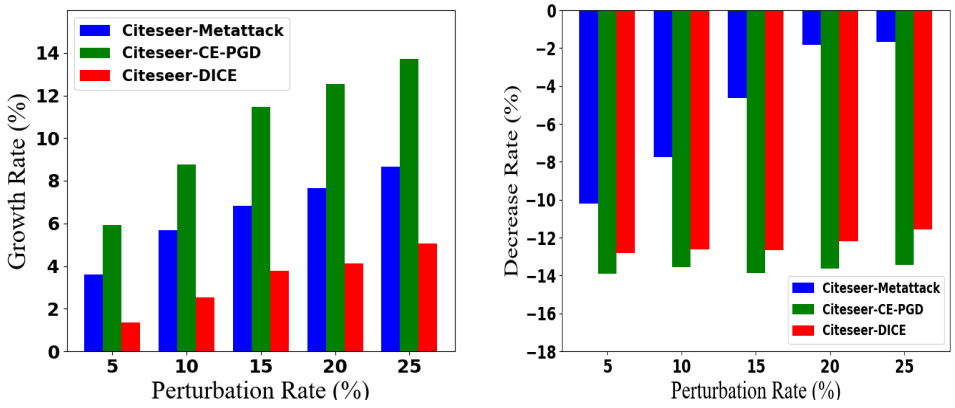
Table 12: Derived ASEP on Polblogsare under Metattack with more attack budget

APR (%)	$\tilde{\beta}$	$\tilde{x}$	APR (%)	$\tilde{\beta}$	$\tilde{x}$
0	81.24	0.5324	**20**	73.36	0.4806
3	79.23	0.5191	**21**	73.26	0.4800
5	78.14	0.5120	**24**	72.78	0.4768
6	77.68	0.5089	**25**	72.70	0.4763
9	76.51	0.5013	**27**	72.31	0.4738
10	76.22	0.4994	**30**	72.31	0.4738
12	75.62	0.4954	**35**	71.92	0.4712
15	74.78	0.4900	**40**	72.50	0.4750
18	74.05	0.4851	**45**	72.48	0.4749



(a) Rank growth under attacks (b) Rank decrease by EquiliRes

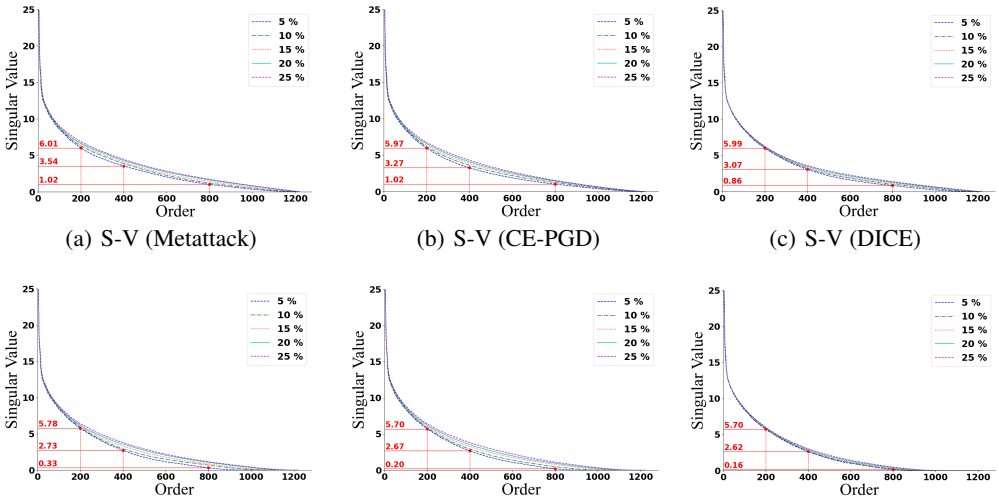
Figure 7: Adjacency matrix of Polblogs.



(a) Rank growth under attacks (b) Rank decrease by EquiliRes

Figure 8: Adjacency matrix of Citeseer.

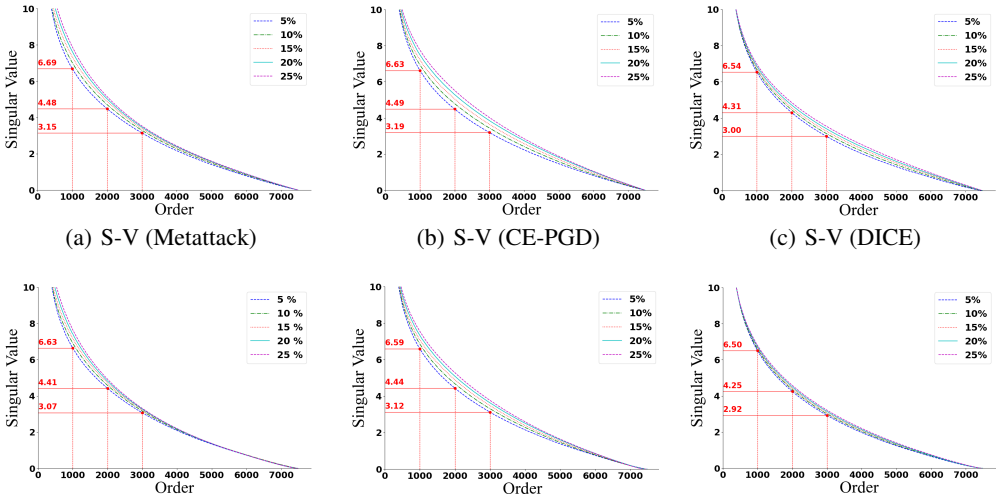
fashion is reasonable and valid to remove adversarial edges w.r.t. different-level adversarial perturbations. In other words, the process of modifying the adjacency matrix to approach to the equilibrium state exactly corresponds to the manipulation of declining singular values. Second, different types of graph adversarial attacks with different attack strengths cause the singular values to have different-level lifts. Correspondingly, our EquiliRes also results in different-level descent in parallel.



(a) S-V (Metattack) (b) S-V (CE-PGD) (c) S-V (DICE)

(d) S-V decrease (Metattack) (e) S-V decrease (CE-PGD) (f) S-V decrease (DICE)

Figure 9: Singular-Value (S-V) variation of Polblogs with EquiliRes.



(a) S-V (Metattack) (b) S-V (CE-PGD) (c) S-V (DICE)

(d) S-V decrease (Metattack) (e) S-V decrease (CE-PGD) (f) S-V decrease (DICE)

Figure 10: Singular-Value (S-V) variation of Amazon Photo with EquiliRes.

J.8 Computation Overhead and Scalability

One advantages of our EquiliRes lies in post-learning that can be obtained from Laplace Transfer without executing deep-learning manipulation in reality, i.e., the critical state of adversarial resilience of graph regime, stemming from the theoretic guarantee of asymptotic stability. Furthermore, our method employs the condensed one-dimensional function to sketch the entire graph' dynamics via degree centrality, rather than other sophisticated multi-dimensional functions, which also decreases the running time to a large extent. To comparatively analyze the efficiency, we calcu-

late the time overhead from the training input to defense output with the experiment settings: APR is set from 0% to 25%, the designated running epoch is set as 200 rounds, and the time overhead is counted on average for 5 times for each experiment. The results on the four datasets Polblogs, Cora_ML, Cora, and Citeseer under the attacks CE-PGD are plotted in Fig. 11.

From the experimental results, we can observe that: i) compared to GAT, HANG, Ours+GAT, and Ours+HANG, the basic GCN, GCN-SVD and our EquiliRes take less time cost on the datasets, this is because GCN only needs to capture the graph structure information without other handling. The singular value decomposition makes GCN-SVD cost similar or a little more overhead than GCN, this is because the decomposition only causes the graph to remove some edges. For our EquiliRes, due to the extra location on equilibrium points, thus it costs a little more than GCN-SVD; ii) Compared with other methods, Mid-GCN and Ours+Mid-GCN consume more time remarkably, we think this is because they need capture the mid-frequency signals from the high-frequency and low-frequency signals through leveraging a filter for graphs, which costs relatively-long time; and iii) although Ours+GAT, Ours+HANG, and Ours+Mid-GCN need a little more overhead than respective individuals, nevertheless, the introduction of our method would not bring too much cost because of the pre-location of inferred equilibrium-point in theory. Overall, these methods are all lightweight compared to those deep learning-based adjacency-matrix iterative optimization mechanisms, such as Pro-GNN-fs (Jin et al., 2020). As shown in Fig. 12, Pro-GNN-fs costs much more time than the others, deriving from the repeatedly search for the optimal low-rank adjacency matrix. For instance, the maximum overhead of Pro-GNN-fs exceeds 300 seconds on dataset Cora_ML, while the maximum of the others is no more than 14 seconds. Comparatively analyzing from the two groups of computation overhead, we know that all these adjacency matrix-based and mid-frequency signal-captured defense approaches cost much less time than the deep learning-based adjacency-matrix iterative optimization methods.

Scalability of Computational Overhead. Given the purification operations of these adversarial defense methods are all edged-centered, thus, we calculate the computational overhead as the number of edges enlarges regarding the four scalable datasets, and the experimental results are listed in Table 13, from which we can see that the proposed EquiliRes is lightweight like GCN and GCN-SVD, compared to other baselines like Hang, Mid-GCN.

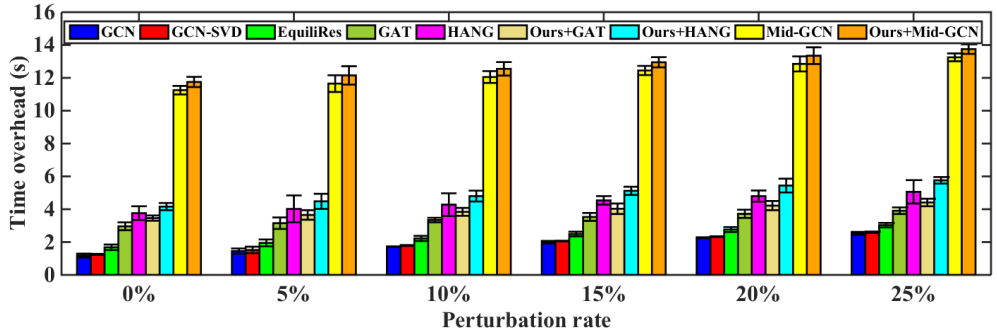
Table 13: Time overhead (s) at the rate of adversarial perturbation 15% (Ns, Es respectively denote Nodes and Edges)

Dataset	**Citeseer (2110Ns 3668Es)**	**Cora (2485Ns 5069Es)**	**Cora_ML (2810Ns 7981Es)**	**Polblogs (1222Ns 16714Es)**
GCN	1.20±0.09	1.18±0.01	1.22±0.04	1.37±0.06
GCN-SVD	1.22±0.02	1.29±0.05	1.50±0.11	1.42±0.01
EquiliRes	1.80±0.14	1.65±0.18	1.74±0.11	1.87±0.12
HANG	3.50±0.29	3.77±0.46	3.95±0.35	4.06±0.24
Ours+HANG	10.70±0.28	4.64±0.37	4.43±0.38	4.51±0.23
GAT	3.30±0.27	3.59±0.19	3.93±0.09	3.31±0.22
Ours+GAT	10.40±0.35	4.46±0.36	4.46±0.42	3.66±0.30
Mid-GCN	10.28±0.32	10.91±0.15	10.58±0.23	11.75±0.26
Ours+Mid-GCN	10.30±0.36	11.45±0.15	11.08±0.25	12.25±0.29

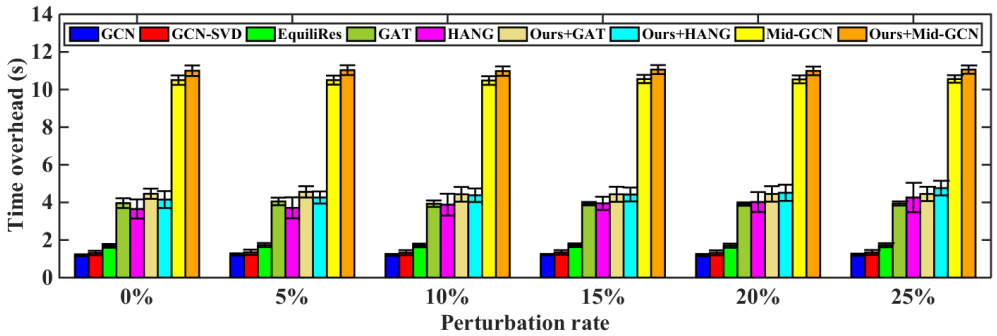
K Related Work

K.1 Graph Adversarial Attack

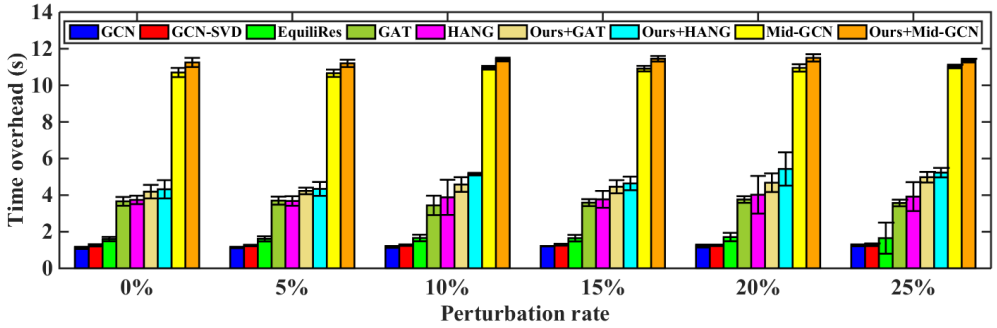
Currently, extensive research has witnessed the graph data is suffering from severe adversarial attacks (Wang & Gong, 2019; Liu et al., 2019; Bojchevski & Günnemann, 2019), partially deriving from the sophisticated dependency (links) among graph nodes. For instance, from the viewpoint of global graph, Wang & Gong (2019) manipulates graph structure towards the collective classification question. Metattack (Zügner & Günnemann, 2019) leverages meta-learning to strategically attack



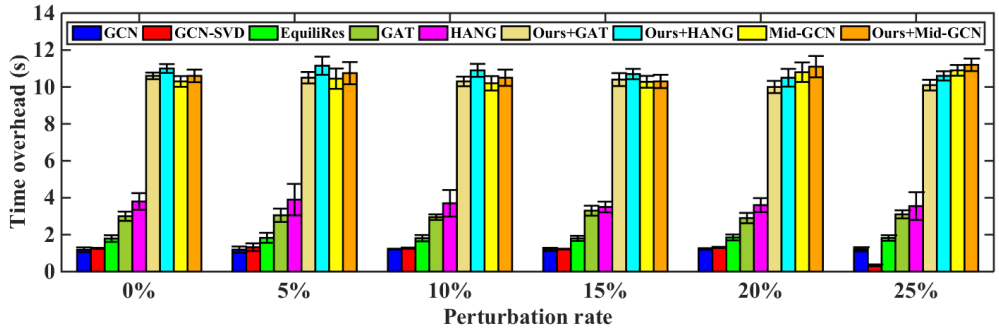
(a) Polblogs



(b) Cora_ML

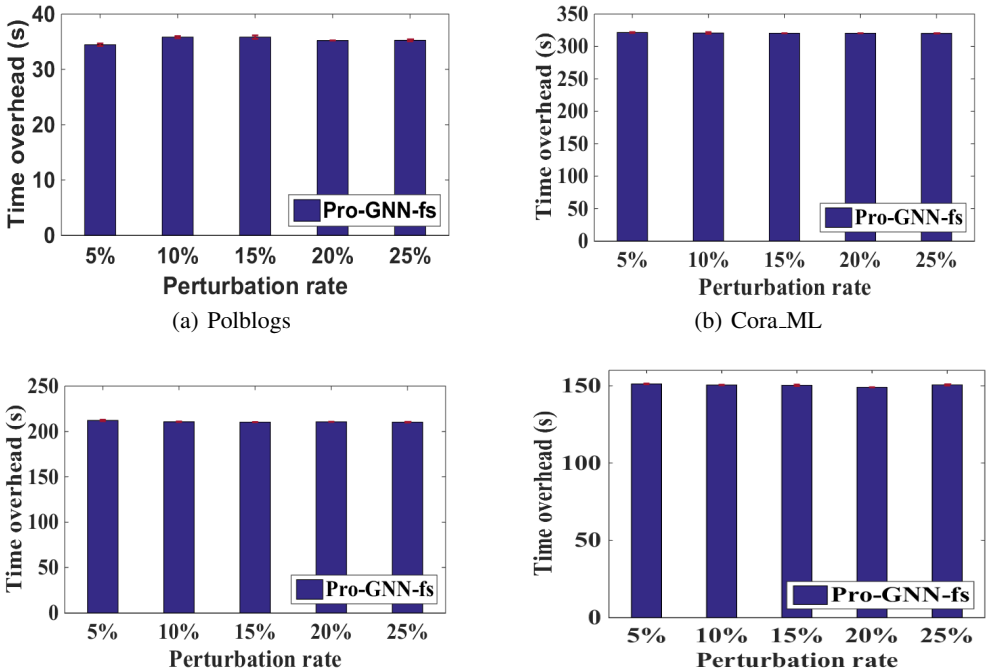


(c) Cora



(d) Citeseer

Figure 11: Computation overhead under CE-PGD.



(a) Polblogs (b) Cora_ML

(c) Cora (d) Citeseer

Figure 12: Time overhead of Pro-GNN-fs under CE-PGD.

the edges, in addition to modify weight, and Zügner & Günnemann (2019) reported that Metattack could quickly boost the rank of adjacency matrix, and tended to establish connections between dissimilar nodes. Xu et al. (2019) employ negative cross entropy-based PGD (CE-PGD) and CW attack-based PGD (CW-PGD) to generate topology perturbations. From the angle of network community, DICE (Waniek et al., 2016) launches an attack on social networks through deliberately removing internal connectivity and building external connectivity to degrade the effectiveness of node classification. Recently, GOttack (Alom et al., 2025) aims to target a specific node in a targeted, structural, direct poisoning attack to alter its predicted class, from the perspective of topology.

K.2 Graph Adversarial Defense

At present, the adversarial defense is studied mainly from two lines: graph per se and graph neural flow. The former mainly improves the adversarial resilience through modifying topology and node feature, for example, from the preprocessing perspective, Wu et al. (2019) firstly remove the dissimilar edges resorting to node feature-based Jaccard-similarity comparison with a preset threshold. GCN-SVD (Chamberlain et al., 2021b), serving as a preprocessing means as well, utilizes SVD to decompose adjacency matrix of the perturbed graph and acquires a low-rank adjacency-matrix approximation to represent a cleaner graph. Pro-GCN-fs (Jin et al., 2020) tries to learn a low-rank and close adjacency matrix to the original. Obviously, the above methods all aim to modify the graph topology and/or features to promote adversarial resilience. Our work falls into this category due to the consideration of finding out the intrinsic critical state of adversarial resilience through only modifying graph topology, without considering node feature.

The other line, i.e. neural ODE network, aims to improve the adversarial resilience by studying the graph neural flow (GNF), that is, neural network constrains the input and output to adhere to certain physical laws, and the transmission of such constrained data within the neural network is referred to as GNF. Typically, these constraints are implemented through system equations from physics. In other words, GNNs can be instantiated with system equations to learn graph data. Recently, the neural ODE has been successfully applied to GNN through modeling information exchange. The studies Chamberlain et al. (2021b); Thorpe et al. (2022) deem message-passing process as the

heat diffusion model, while the references Chamberlain et al. (2021a); Song et al. (2022) regard the message-passing process as Beltrami diffusion. In addition, towards graph nodes, Rusch et al. (2022) adopt the oscillators to model nodes and conduct the message-passing process under the guidance of a coupled ODE. Recently, HANG (Zhao et al., 2023) imposes an energy-conservation constraint on GNF to enhance the adversarial robustness of GNNs. Upon the incorporation of OOD detection, GOOD-AT (Li et al., 2024) leverages the popular adversarial training paradigm to defend against poisoning and evasion attacks. Differently, our approach models graph perturbations by adhering to the asymptotic-stability theory, and proposes a preprocessing optimization method tailored for graph data through resorting to asymptotically-stable equilibrium-point over topology.

L Motivation and Open Issues

Motivation. Towards why we need dynamic-system findings to study adversarial learning, We think there are at least three reasons: i) the oscillation in dynamic system and the perturbation in adversarial learning have common characteristic-dynamics. Considering the difficulty of measuring adversarial attacks (To our best knowledge, no solution to quantify adversarial attacks to date), we can employ *Laplace Transform* to measure adversarial attack quantitatively, especially for the scenarios wherein we have no idea about the behavior of GAAs. Hence, resorting to the findings in dynamic system to bridge connection to adversarial learning is proper and worth in-depth exploration; ii) using the stability theory in dynamic systems, we can infer an optimal robust (equilibrium-point) state in theory for each graph regime in a simple yet valid manner; and iii) in the light of equilibrium point, we can directly purify the perturbed adjacency matrix in a light-weight way, resulting in a lower computational overhead compared to the time-consuming model-retraining methods (e.g. Pro-GNN-fs).

Open Issues. Our work is completely a new attempt to study graph's intrinsic adversarial resilience under the guidance of equilibrium theory, and three key issues are worth discussing deeply: i) **Condensed One-Dimensional Mapping Function.** In our work, a simple condensed one-dimensional function is developed to map the dynamics of graph as a whole, in fact, two-dimensional or higher dimensional mapping can be explored as well for distinct graph regimes, this is because our proposed asymptotic-stability theoretic framework fully supports multi-dimensional function; and ii) **Adversarial Resilience on Multi-Modal Data.** This paper currently concentrates graphs, a natural question emerges, i.e. whether other modal data also has such equilibrium state to keep attack-resilient, such as image, audio, text, etc. If exist, whether we can tackle in the same way in technique, regarding totally different data structure; and iii) **Parameters of Adversarial Resilience-Declining Function.** Different graph regimes usually have distinct adversarial resilience-declining tendencies under different-level adversarial perturbations.