Stronger and Transferable Node Injection Attacks

Samyak Jain and Tanima Dutta
Indian Institute of Technology (BHU) Varanasi
samykjain.cse18@itbhu.ac.in, tanima.cse@itbhu.ac.in

Abstract
Despite the increasing popularity of graph neural networks (GNNs), the security risks associated with their deployment have not been well explored. Existing works follow the standard adversarial attacks to maximize cross-entropy loss within an L-infinity norm bound. We analyze the robustness of GNNs against node injection attacks (NIAs) in black-box settings by allowing new nodes to be injected and attacked. In this work, we propose to design stronger and transferable NIAs. First, we propose margin aware attack (MAA) that uses a maximum margin loss to generate NIAs. We then propose a novel margin and direction aware attack (MDA) that diversifies the initial directions of MAA attack by minimizing the cosine similarity of the injected nodes with respect to their respective random initialization in addition to the maximization of max-margin loss. This makes the NIAs stronger. We further observe that using L-infinity norm of gradients in the attack step leads to an enhanced diversity amongst the node features, thereby further enhancing the strength of the attack. We incorporate transferability in NIAs by perturbing the surrogate model before generating the attack. An analysis of Eigen Spectrum Density of the hessian of the loss emphasizes that perturbing the weights of the surrogate model improves the transferability. Our experimental results demonstrate that the proposed resilient node injection attack (R-NIA) is significantly stronger and transferable than existing NIAs on graph robustness benchmarks.

Introduction
Graph neural networks (GNNs) have gained success in performing learning and inference on graph-structured data (Kipf and Welling 2017b; Hamilton, Ying, and Leskovec 2017; Zhou et al. 2020; Ye et al. 2022) An increasing amount of attention is being paid on node injection attacks (NIAs) (Chen et al. 2022; Zou et al. 2021; Wang and Gong 2019; Sun et al. 2020) that aim to insert new nodes into the graph without making changes in the existing nodes or structure. An attacker can make new accounts in the social networking graph and manipulate the attributes of those accounts in a way such that original nodes get fooled and suggest wrong recommendations to existing users. Similarly, in the case of graphs on financial data where the transactions between the customers and merchants are stored in a graph, fooling the fraud detection system by making fake accounts can lead to catastrophic results. Gradient-based optimization can be done to perturb the node features and identify the appropriate locations to inject new nodes. One of the popular attacks, projected gradient descent (PGD) attack (Madry et al. 2018), proposes to maximize cross-entropy loss to generate the attack. Carlini and Wagner (2017); Gowal et al. (2019) show that max-margin loss can generate stronger attacks on images as compared to PGD (Madry et al. 2018).

As shown in Figure 1 the white dot ball represents the threat model in which the attack is constrained to remain. Within the defined threat model, the attacker can fool the model to classify class C3. Since PGD only minimizes the probability of the true class and does not take the probability of the true target class (class C3), as shown in Figure 1 (a), on perturbing the node features using PGD, it can end up traversing along the loss contour lines of class C3. Thus, the loss with respect to class C3 remains almost constant. However, maximizing the max-margin loss remains almost constant. However, maximizing the max-margin loss leads to traversal in the direction orthogonal to the loss contour lines (Figure 1 (b)). Motivated by this, we propose to maximize max-margin loss instead of cross-entropy. We term this attack as margin aware...
We show that the new loss landscape generalizes better in we elaborated the proposed attacks. We finally show the works well for adversarially trained GNNs. Related works and preliminaries in next two sections. Then, we also show graphs like Aminer, Amazon2M, and Twitter. We also show • that perturbing the surrogate model within an \( \ell_2 \) norm ball in the weight space before initiating the attack leads to improved transferability of the attack generated on the surrogate model. This may happened because the attack is no longer specific to the optimal solution of the surrogate model and therefore generalizes better to other models. We summarize our contributions as follows:

• PGD only minimizes the probability of the true class and does not take the probability of the true target class into account while generating the node injection attack can enhance diversity amongst the attacked features. We also find that perturbing the surrogate model within an \( \ell_2 \) norm ball in the weight space before initiating the attack leads to improved transferability of the attack generated on the surrogate model. This may happened because the attack is no longer specific to the optimal solution of the surrogate model and therefore generalizes better to other models. We summarize our contributions as follows:

• We incorporate directionality in MAA by minimizing the cosine similarity with respect to the random initial direction in initial attack iterations to better explore the attack constraint space. We named the model as margin and direction aware attack (MDA) that explores the attribute space and result in perturbations that are diverse and exploratory.

• We demonstrate the use of \( \ell_2 \) norm gradient ascent, instead of the commonly used \( \ell_\infty \) norm, that leads to an improved attack strength, since there is no constraint of variation in the features of the injected nodes.

• We propose to improve the attack transferability by perturbing the weights of the surrogate model within an \( \ell_2 \) norm constraint in the weight space, before generating the attack. We show that the new loss landscape generalizes better in a black box attack setting. We term this attack as margin, direction and transferability aware attack (MDTA).

While MetaAttack (Zügner et al. 2020) and NAttack (Zügner, Akbarnejad, and Gümennemann 2018) were originally proposed as poisoning attacks, they have been adopted for node injections in Wang et al. (2020). As shown in Wang et al. (2020), the adaptation of MetaAttack and NAttack outperforms FGSM (Szegedy et al. 2013), but it is computationally very expensive. Motivated by this, Wang et al. (2020) propose Approximate Fast Gradient Sign Method (AFSM), which performs similarly but is computationally much cheaper than MetaAttack and NAttack. In this work, we compare the proposed method with MetaAttack, NAttack and AFGSM as well. Recently, Chen et al. (2022) introduced a regularizer based on cosine similarity, which ensures that the homophily between the features of the added nodes and the original nodes in the graph is maintained. This helps in the generation of imperceptible attacks while being strong. Since it is important to testify that the proposed attack remains imperceptible in nature, we perform a study on the imperceptibility of the proposed method and utilize the closest attribute distance (CAD) (Zou et al. 2021) as the metric. While imperceptibility is important but it is also important to ensure that the attack remains strong. We demonstrate that the proposed attack, while being imperceptible, also outperforms the existing methods on the Graph robustness benchmark (Zheng et al. 2021).

**Preliminaries**

**Threat Model:** We consider the same threat model, as considered in Chen et al. (2022). The aim of the attack is to fool a GNN \( f_0 \) for graph \( G = (A, X) \) by generating another graph \( G' = (A', X') \), where \( A \in \mathbb{R}^{d \times d} \) denotes the graph adjacency matrix and \( X \in \mathbb{R}^{d \times b} \) denotes the node feature matrix. Here, \( d \) is the number of nodes in the original graph and each node has a feature vector of dimension \( b \). The number of new nodes injected is given by \( n \). The objective function is:

\[
\max_{\|G' - G\| \leq \Delta} \ell_{\text{atk}}(f_0(G')),
\]

\[
\Delta \rightarrow \{ |X_{\text{atk}}| \leq P \in \mathbb{Z}; |A' - A| \leq R \in \mathbb{Z}; \|X'\|_{\infty} \leq \epsilon \in \mathbb{R}\}.
\]
where $\Delta$ is the constraint on the perturbed graph (afore-said three constraints) and $\ell_{atk}$ is the attack loss which is maximized in order to fool the GNN. $A' = [A; A_{atk}]$, $X' = [X; X_{atk}]$ and ` denotes concatenation operation. $\{P, R\}$ are the upper bound on the number of injected nodes and the number of new edges inserted, respectively. The range of new nodes injected is also bounded. The loss is perturbed on a given input by limiting the upper value of features by $\epsilon$.

**Black Box Setting:** In black box setting, any information about the architecture of the system under attack is kept secret. Similar to (Lord, Mueller, and Bertinetto 2022; Chen et al. 2022), we train a surrogate GNN to check the transferability of the attack. We test the attack transferability for same and different architectures of surrogate and black box models.

### Rethinking Strength of Node Injection Attacks

**Margin-aware Node Injection Attack (MAA)**

The maximization of the cross-entropy loss in PGD attack (Madry et al. 2018) leads to the minimization of the originally predicted class confidence only since the cross-entropy loss in PGD does not consider the confidence of other classes. Figure 1 (a) shows that decreasing the confidence of one class need not decrease the confidence of the potential attack class (class whose boundary is closest to input) in a multi-classification setting, and this results in no change in prediction. On the contrary, it is observed in Figure 1 (b) and Figure 1 (Supplementary) that by maximizing the max-margin loss, the above scenario can be avoided. The max-margin loss $\ell_{mm}$ is given by:

$$\ell_{mm}(f_\theta) = -f^y_\theta(x) + \max_{i \neq y} f^i_\theta(x). \quad (2)$$

Thus maximizing the max-margin loss leads to the maximization of the margin between the true class and the highest confident false class via decreasing the confidence of the true and maximizing the confidence of the false class. Motivated by this, we propose margin aware attack (MAA), which maximizes max-margin loss to generate the attack. We empirically show in Table 1 that MAA attack is stronger than PGD (Madry et al. 2018) by over 1.3% on Citeseer (Giles, Bollacker, and Lawrence 1998) dataset.

**Leveraging Directional Similarity in MAA (MDA)**

Max-margin loss helps in achieving a larger norm between original and attacked node features. However, Figure 1 (b) depicts that it is biased towards the smoothness of the local space near the initialization point. Since local smoothness need not lead to the best trajectory globally, thus simply maximizing max-margin loss can lead to suboptimal attack generation. Motivated by this, we propose:

$$\ell_{mda}(f_\theta(X \oplus X_{atk})) = \ell_{mm}(f_\theta(X \oplus X_{atk})) - \gamma \times \ell_{co-sim}(f_\theta(X \oplus X_{init}), f_\theta(X \oplus X_{atk})), \quad (3)$$

where $X_{init}$ denotes the randomly initialized feature matrix of the injected nodes. $X_{atk}$ denotes the feature matrix of the injected nodes at some time stamp of the attack generation. $\oplus$ denotes the concatenation operation. The max-margin loss ($\ell_{mm}$) is defined in Eq. (2) and the cosine similarity loss $\ell_{co-sim}$ is defined as follows:

$$\ell_{co-sim}(f_\theta(X \oplus X_{init}), f_\theta(X \oplus X_{atk})) = \frac{f_\theta(X \oplus X_{init})^T f_\theta(X \oplus X_{atk})}{||f_\theta(X \oplus X_{init})|| \cdot ||f_\theta(X \oplus X_{atk})||}. \quad (4)$$

We further present a detailed explanation of the proposed MDA attack using a toy example in Supplementary.

### Rethinking $\ell_\infty$ Norm for Gradient Ascent

Past works Madry et al. (2018); Chen et al. (2022) naively use the sign of the gradients for crafting attacks. The sign of the gradient (in case of $\ell_\infty$ attack) in the attack generation...
Figure 4: Eigenvalues spectrum of the loss calculated on training set. A: smaller eigenvalues have similar density. B: dominant larger eigenvalues have different densities. This indicates two loss landscapes are very different.

As demonstrated by Lord, Mueller, and Bertinetto (2022), model is computed at points \( \theta \) perturbation bound of over-fitted to its minima. Now, if a perturbation within the rep-resentation of the loss curve for the surrogate model. Let \( \ell_2 \) norm of the gradients for the black box model and the green curve get, perturbed weights as follows: to perturb the weights of \( GNN_i \) tures. Let, the landscape that is more generalizable across different architec-tures. As shown in Table 1 and Table 2 in Supplementary, this observation leads to improvements upto 7% on \( \ell_\infty \) norm.

Enhancing Transferability in MDA (MDTA)

As demonstrated by Lord, Mueller, and Bertinetto (2022), merely generating strong attacks on surrogate models does not guarantee strong attack transfer, thus leading to poor generalization. On the other hand, Foret et al. (2021) showed that flatter minima lead to improved generalization. Motivated by this, we analyze the effect of generating attack on a perturbed surrogate model so that the attack generated is on a loss landscape that is more generalizable across different architectures. Let, the \( i^{th} \) node feature vector \( x_i \) and corresponding ground truth label \( y_i \), on maximizing the cross-entropy loss to perturb the weights of GNN \( f \), parameterized by \( \Theta \), we get, perturbed weights as follows:

\[
\tilde{\Theta} = \arg\max_{\Theta \in \mathcal{M}(\Theta)} \frac{1}{n} \sum_{i=1}^{n} \ell_{ce}(f_\Theta(x_i), y_i). (5)
\]

To develop a better understanding, we present a motivational example in Figure 2, where the black curve represents the loss curve for the black box model and the green curve represents the loss curve for the surrogate model. Let \( \theta_0 \) represent the trained weights of the surrogate model, which have over-fitted to its minima. Now, if a perturbation within the perturbation bound of \( \ell_2 \) norm radius \( ||\theta_0 - \theta'_0|| \) is generated, then the weights of the models \( (\theta_0) \) can change to \( \theta'_0 \) after addition of the perturbation. If the loss of the black box model is computed at points \( \theta_0(L_1) \) and \( \theta_0(L_2) \), then it is clear that \( L_2 < L_1 \). Therefore, perturbing the weights of the surrogate model can help in finding the weights which have better local properties in the black box model. Therefore, generating an attack from the perturbed surrogate model can lead to improved transferability. Therefore, we propose that:

We show the Eigen Spectrum Density of Hessian of the loss calculated using the black box and surrogate models with the same GCN architecture and trained independently in Figure 4. Here, the spectral density of black box and surrogate models is similar for small eigenvalues and quite different for larger eigenvalues which are responsible for determining the sharpness of the loss landscape (Foret et al. 2021). Therefore, perturbing the model in \( \ell_2 \) norm ball can help in matching the spectral density of larger eigenvalues, thus improving the transferability. We quantify the similarity between the spectral distributions for the black box and perturbed surrogate models in Figure 3. We use cosine similarity and KL Divergence as the similarity metrics to compare the two distributions. It is observed that perturbing the surrogate model using low to moderate perturbation bound radius leads to improved similarity. This indicates that perturbing the surrogate model can indeed lead to better alignment between the loss landscapes of the surrogate and black box models. To investigate whether this alignment should lead to increased attack transferability, we analyze the robust accuracy (%) using different \( \ell_2 \) norm and perturbation radius \( (\rho) \) in Table-3 in Supplementary for the Cora dataset. We observe improved transferability by over 3% as the value of \( \rho \) is increased to 0.2. This shows that perturbing the surrogate model can help in improving transferability. Further, since the accuracy of the surrogate model itself drops significantly at larger values of \( \rho \), we observe poor transferability.

Resilient Node Injection Attack (R-NIA)

We propose resilient node injection attack (R-NIA), which is described in Algorithm 1. Given a black box model \( g_\Theta \), we first train a surrogate model \( f_\Theta \). Then, an attack is performed using this model and the black box \( g_\Theta \) is used for evaluating the attack. After randomly initializing the attacked adjacency matrix and the attacked feature matrix (Line 3) with gaussian distribution, we first maximize the standard cross-entropy loss (Zhang et al. 2019; Wu, Wang, and Xia 2020), in order to come out of the sharp minima of the surrogate model where it might have converged to. For this, we use a \( \ell_2 \) norm with a bound of \( \rho = 0.2 \). The details are presented in Line 4. The detailed study on the effect of the value of \( \rho \) is discussed in Table 3 in Supplementary. After perturbing the model, we get a new model \( f_\Theta \). In order to identify the right locations where new nodes should be injected (similar to Chen et al. (Chen et al. 2022)), we adopt a gradient-based attack on the adjacency matrix. We calculate the gradients on the adjacency matrix and take the \( \ell_\infty \) norm of the gradients.
Algorithm 1: Resilient Node Injection Attack (R-NIA)

1: **Input**: $f_\theta$, $G = ((A, X))$;
2: **Output**: $A_{atk}$, perturbed weights $\tilde{\Theta}$;
3: Randomly initialize the matrix $(A_{atk}, X_{atk})$; $A_{atk} = 0.1 \cdot N(0, 1)_{n \times n}$, $X_{atk} = X_{init}$ and $G' = (A \oplus A_{atk}, X \oplus X_{atk})$;
4: $\tilde{\Theta} = \arg\max_{\Theta \in \mathcal{M}(\Theta)} \sum_{i=1}^{d} \ell_{ce}(f_\Theta(x_i), y_i)$; \% $\ell_{ce}$ denotes the standard cross-entropy loss. \%
5: $A_{atk} = A_{atk} + \text{sign} (\nabla A_{atk} \times \ell_{ce}(f_\Theta(X \oplus X_{atk})))$;
6: $A'_{atk} = \text{topk}(A_{atk})$;
7: $A_{atk} = \text{round}(A'_{atk}, 0, 1)$;
8: $\ell_{mda}(f_\Theta(X \oplus X_{atk})) = \ell_{mon}(f_\Theta(X \oplus X_{atk})) - \gamma \times \ell_{cox-sim}(f_\Theta(X \oplus X_{init}), f_\Theta(X \oplus X_{atk}));$
9: if $\text{iter} >= \frac{2}{3} I$ then \%
10: end if
11: $X_{atk} = X_{atk} + \alpha \times \frac{(\nabla X_{atk} \cdot \ell_{mda}(f_\Theta(X \oplus X_{atk}))}{|| (\nabla X_{atk} \cdot \ell_{mda}(f_\Theta(X \oplus X_{atk})) ||}$;
12: $X_{atk} = \text{clamp}(X_{atk}, \text{MAX}, \text{MIN})$; \% MIN and MAX depend on range of original nodes. \%

$G' = (A \oplus A_{atk}, X \oplus X_{atk})$.

We present the experimental results of PGD (Madry et al. 2018), MAA (Margin Aware Attack), MTA (Margin and Transferability Aware attack), MDA (Margin and Direction Aware attack), MDTA (Margin, Direction And Transferability Aware attack), and R-NIA (Resilient Node Injection Attack) with $\ell_2$ (R-NIA) and $\ell_{\infty}$ norm (R-NIA-$\ell_{\infty}$) of gradients in Table 1 for Cora and Citeseer datasets and Table 2 in Supplementary for Flickr dataset. Further results on large datasets, like Aminer, Amazon2M, and Twitter, are present in Table 2. We utilize PGD (Madry et al. 2018), PGD with Harmonious Adversarial Objective (HAO) (Chen et al. 2022), TDGIA (Zou et al. 2021), MetaAttack (Wang et al. 2020), G-NIA (Tao et al. 2021) and Approximate FGSM (AFGSM) (Wang et al. 2020) as baselines. We consider two variants of MetaAttack. Either adding all the nodes at once (MetaAttack (one time)) or adding them in a sequential manner (MetaAttack (sequential)). G-NIA (Tao et al. 2021) aims to craft a single node which can fool certain target nodes of the graph. For a fair comparison, we modified the optimization problem of (Tao et al. 2021), to maximize the loss for all the injected nodes in the graph and attack all the original nodes.

**Evaluation Results**

We perform the experimental results of PGD (Madry et al. 2018), MAA (Margin Aware Attack), MTA (Margin and Transferability Aware attack), MDA (Margin and Direction Aware attack), MDTA (Margin, Direction And Transferability Aware attack), and R-NIA (Resilient Node Injection Attack) with $\ell_2$ (R-NIA) and $\ell_{\infty}$ norm (R-NIA-$\ell_{\infty}$) of gradients in Table 1 for Cora and Citeseer datasets and Table 2 in Supplementary for Flickr dataset. Further results on large datasets, like Aminer, Amazon2M, and Twitter, are present in Table 2. We utilize PGD (Madry et al. 2018), PGD with Harmonious Adversarial Objective (HAO) (Chen et al. 2022), TDGIA (Zou et al. 2021), MetaAttack (Wang et al. 2020), G-NIA (Tao et al. 2021) and Approximate FGSM (AFGSM) (Wang et al. 2020) as baselines. We consider two variants of MetaAttack. Either adding all the nodes at once (MetaAttack (one time)) or adding them in a sequential manner (MetaAttack (sequential)). G-NIA (Tao et al. 2021) aims to craft a single node which can fool certain target nodes of the graph. For a fair comparison, we modified the optimization problem of (Tao et al. 2021), to maximize the loss for all the injected nodes in the graph and attack all the original nodes.

**Is using $\ell_{\infty}$ norm of gradients (Madry et al. 2018; Chen et al. 2022) the best choice?** As seen in the bottom half of Table 1 and Table 2 (Supplementary), we find that using $\ell_2$ norm of the gradients instead of $\ell_{\infty}$ leads to a significant drop in the classification accuracy. Due to enhanced diversity in the features in the case of $\ell_2$ norm, we observe stronger attack with upto 7% higher attack strength.

**Is maximizing Cross-Entropy loss a good choice in PGD attack formulation?** We compare the results of PGD to update the adjacency matrix $(A_{atk})$, as shown in Line 5. In order to follow the constraints on the number of inserted edges, we take the topk values from the adjacency matrix and make others as zero, where e denotes the upper bound on the number of new injected edges. This gives a new adjacency matrix $A'_{atk}$ (Line 6). Finally, rounding off is taken on the new adjacency matrix $A'_{atk}$ (Line 7). We use a single-step gradient ascent to generate $A'_{atk}$ as the adjacency matrix is unable to take continuous values. We further show an ablation on varying the number of attack steps to attack the adjacency matrix in Figure 5(a) in Supplementary.

We modify the injected node features so that the nodes belonging to the original graph $G$ getfooled. For this, we utilize a combination of max-margin and cosine similarity loss (Line 8). We perform an iterative gradient ascent using the proposed MDA loss with $I$ number of iterations and step size of $\varepsilon$, where we maximize the max-margin loss while minimizing the cosine similarity loss between the random initialization and the attacked features of the injected nodes. We minimize the cosine similarity only for 2/3 of the total number of iterations $I$ (where $I = 1000$), as shown in Line 10. As discussed previously, using $\ell_2$ bound on gradients helps in enhancing the diversity and therefore leads to stronger attacks. Thus, we propose to take $\ell_2$ norm of the gradients in Line 11. Finally, clipping is performed (Line 12) to ensure that constraints are maintained. At inference time, instead of the surrogate model, the black box model $g_{\Theta}$ is used for evaluation on the perturbed graph $G'$. 

**Experiments Results**

**Training Details**

We empirically test our models on six datasets, i.e., Cora (Yang, Cohen, and Salakhutdinov 2016), Cite-
and MAA in Table 1 and Table 2 (Supplementary). We observe that the performance of MAA is better than PGD by over 1.5% on Cora and Flickr and 2% on the Citeseer.

★ Can maximizing max-margin loss alone lead to sub-optimal search in output constraint space? Based on Table 1 and Table 2 (Supplementary), it is clear that MAA leads to stronger attacks than PGD. However, minimizing cosine similarity in addition to MAA improves attack strength by up to 12% on Cora (MDA, MDA-\(\ell_2\)) in comparison to PGD. This shows the importance of incorporating directional similarity in the attack objective. The attack optimization is biased on local smoothness of loss landscape. Minimizing cosine similarity helps in better optimization of the attack.

Based on the results, it is important to analyze the effect of minimizing cosine similarity without maximizing max-margin loss in MDA. It is clear from the comparison between PGD (Co-Sim) and PGD that simply minimizing the cosine similarity does not lead to strong attacks. This is because, unlike max-margin loss maximization, minimizing cosine similarity alone does not lead to a particular direction where a class can be changed. But it helps in finding better initialization that can generate strong attacks. Motivated by this, we use cosine similarity for the initial 750 attack iterations out of the total 1000 iterations to find a better starting point.

★ Can we enhance the transferability of the attacks generated on a surrogate model? Based on Table 1 and Table 2 (Supplementary), we observe that it is indeed possible to improve the transferability of the attacks. The comparison of MDA with MDTA clearly shows that the transferability of the attacks for different architectures (CN→AT and AT→CN) has improved by up to 3%. Finally, while we observe that HAO (Chen et al. 2022) is the stronger baseline in Table 1, the proposed R-NIA shows improvements up to 8.18% over it on the Cora and around 5% on the Citeseer datasets.

![Figure 5](attachment:fig5.png)  
**Figure 5:** Impact on transferability between (a) adversarial (A) and standard (S) trained models; (b) non-local GCN (N) and GCN (G) models.

### Impact of Graph Size, #Inject Node & #Inject Edge

As shown in Table 2, we observe that on adding a larger number of nodes and edges the proposed R-NIA attack improves the performance over PGD by over 8.67% on Flickr and by over 10% on an even larger dataset like Aminer. We also observe gains on up to 10 times further larger datasets (larger graph size - Table 1 in Supplementary) like Amazon2M and Twitter datasets. On Amazon2M, we outperform PGD by up to 6.09% and by 7% on Twitter dataset. We observe that as the number of nodes and edges are allowed to be injected, the strength of PGD attack increases relatively slowly as compared to R-NIA attack.

### Transferability on Robust Models

Figure 5 (a) checks the attack transferability where the surrogate model is either an adversarially trained model (A) or a standard trained model (S). We train the GCN and GAT models using ten steps PGD adversarial training to get the robust models (denoted by A). We observe transferability gains between robust models (A→A) over 15% on the Cora dataset. Further, we observe that R-NIA shows improved...
We provide a detailed study of the impact of $\ell_\infty$ well. On larger graphs, more gains over the PGD baseline are observed on injecting larger number of nodes and edges.

(b) depicts the results of transferability experiments on non-

Table 2: Robust Accuracy (%) on injecting different number of new nodes $N_{inj}$ and edges $E_{inj}$ on Flickr and Aminer datasets, i.e., $\{N_{inj}, E_{inj}\}$ pair. K and M indicate $10^K$ and $10^M$. We take $\ell_2$ norm of gradients for the PGD (Madry et al. 2018) baseline as well. On larger graphs, more gains over the PGD baseline are observed on injecting larger number of nodes and edges.

Figure 6: Results on attack success vs CAD, where attack success = 100− robust accuracy on (a) Cora and (b) Citeseer.

transferability as compared to PGD when the attack is generated using S/A and evaluated on A/S, respectively. Figure 5 (b) depicts the results of transferability experiments on non-local GCN (N). We observe that R-NIA performs better than PGD by up to 13.7% on the CORA dataset for N $\rightarrow$ N. These results demonstrate that the R-NIA attack is stronger than PGD not only on standard models but also on robust models and generalizes to different classes of GNNs as well.

Ablation Studies

We provide a detailed study of the impact of $\ell_2$ norm constraint ($\rho$) for weight perturbation in MTA on the Cora dataset in Table 3 (Supplementary). It is evident that there is a certain range of values of $\rho$, which leads to the strongest transferability. $\rho < 0.2$ does not make any significant change where both surrogate and black box models have the same architecture but helps in improving the transferability rate. A larger value of $\rho$ degrades the original model, thus degrading the transferability. We study the effect of changing the constraints on the number of nodes injected in the threat model, as shown in Figures 2 and 3 (Supplementary), where we observe that MDA and R-NIA consistently generate a stronger attack on using a different number of injected nodes. As shown in Figure 5 (a) in Supplementary, increasing the number of steps to generate an attack on $A_{atk}$ does not lead to significant changes in robust accuracy, therefore we propose to use a single-step attack to generate $A_{atk}$. As shown in Figure 5 (b) in Supplementary using a $\gamma$ value close to 0.1 leads to the strongest attack. The plot shows that $\gamma$ is an important hyperparameter for MDAT. As shown in Figure 4 (a) in Supplementary, while PGD leads to a stronger attack for less number of attack steps, it saturates very early. MDA shows improved performance with the increase in the number of attack steps. Figure 4 (b) in Supplementary shows that MDA outperforms PGD on increasing the number of injected edges.

Imprecipitability Study

It is important to ensure that after node injection attack, the graph still remains imperceptible (Chen et al. 2022; Tao et al. 2022). For this, we use the closest attribute distance (CAD) as the metric (Tao et al. 2022; Zou et al. 2021). For each injected node, CAD calculates the nearest node feature with the smallest $\ell_2$ norm feature distance with the injected node and averages it over all the injected nodes. The results are shown in Figure 6, where we observe that the proposed methods have a lower value of CAD when compared to PGD. Further, R-NIA leads to stronger attacks with a small increase in CAD when compared to HAO. This shows that the proposed methods ensure that the graph is imperceptible after the node injection attacks. As shown by Tao et al. (2022), there is a strong correlation between CAD and two other imperceptibility metrics: Smoothness (Chen et al. 2022) and Graph FD (Tao et al. 2022). Therefore, CAD is a reliable metric.

Conclusion

We highlight the need to rethink node injection attacks on GNNs. To enhance the attack transferability, we propose to perform a weight perturbation in $\ell_2$ norm before crafting the attack. We show that the use of $\ell_\infty$ norm restricts the diversity of attack features and propose to use $\ell_2$ norm instead. We demonstrate improved performance over graph robustness benchmark (Zheng et al. 2021) models. We show that our method is generalizable to larger graphs as well. It improves attack strength in the case of adversarially trained models.

Acknowledgements

We thank SERB MTR/2021/604 and SIR/2022/000536.
References


Kipf, T. N.; and Welling, M. 2017b. Semi-Supervised Classification with Graph Convolutional Networks. In International Conference on Learning Representations.


