SINGLE-NODE ATTACK FOR FOOLING GRAPH NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Graph neural networks (GNNs) have shown broad applicability in a variety of domains. Some of these domains, such as social networks and product recommendations, are fertile ground for malicious users and behavior. In this paper, we show that GNNs are vulnerable to the extremely limited scenario of a single-node adversarial example, where the node cannot be picked by the attacker. That is, an attacker can force the GNN to classify any target node to a chosen label by only slightly perturbing another *single arbitrary node* in the graph, even when *not being able to pick that specific attacker node*. When the adversary is allowed to *pick a specific attacker node*, the attack is even more effective. We show that this attack is effective across various GNN types, across a variety of real-world datasets, and as a targeted and a non-targeted attack.

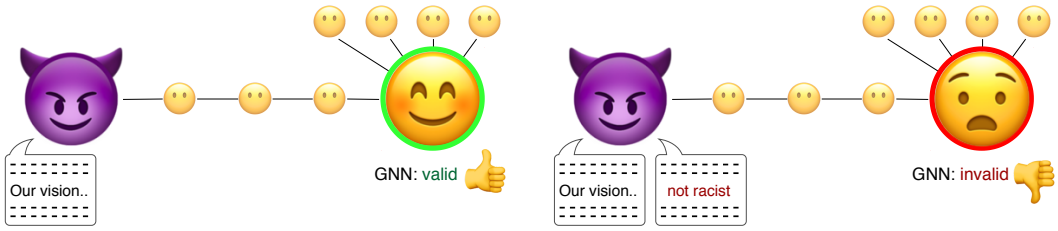
1 INTRODUCTION

Graph neural networks (GNNs) (Scarselli et al., 2008; Micheli, 2009) have recently shown sharply increasing popularity due to their generality and computation-efficiency (Duvenaud et al., 2015; Li et al., 2016; Kipf & Welling, 2017; Hamilton et al., 2017; Veličković et al., 2018; Xu et al., 2019b). Graph-structured data underlie a plethora of domains such as citation networks (Sen et al., 2008), social networks (Leskovec & Mcauley, 2012; Ribeiro et al., 2017; 2018), knowledge graphs (Wang et al., 2018; Trivedi et al., 2017; Schlichtkrull et al., 2018), and product recommendations (Shchur et al., 2018). Therefore, GNNs are applicable for a variety of real-world structured data.

While most work in this field has focused on improving the accuracy of GNNs and applying them to a growing number of domains, only a few past works have explored the vulnerability of GNNs to adversarial examples. Consider the following scenario: a malicious user joins a social network such as Twitter or Facebook. The malicious user mocks the behavior of a benign user, establishes connections with other users, and submits benign posts. After some time, the user submits a new adversarially crafted post, which might seem irregular but overall benign. This new post perturbs the representation of the user as seen by a GNN. As a result, another, specific benign user gets blocked from the network; alternatively, another malicious user submits a hateful post – but does not get blocked. This scenario is illustrated in Figure 1. In this paper, we show the feasibility of such a troublesome scenario: a single attacker node can perturb its own representation, such that another node will be misclassified as a label of the attacker's choice.

Most previous work on adversarial examples in GNNs required the perturbation to span *multiple* nodes, which in reality requires the cooperation of multiple attackers. For example, the pioneering work of Zügner et al. (2018) perturbed a *set* of attacker nodes; Bojchevski & Günnemann (2019a) perturb edges that are covered by a *set* of nodes. To the best of our knowledge, our work is the first to show that a *single*-node attack is as effective as a multiple-node attack. Further and in contrast with existing work, we show that perturbing a single *node* is more harmful than perturbing a single *edge*.

In this paper, we present a first a single-node adversarial attack on graph neural networks. If the adversary is allowed to choose the attacker node, for example, by hacking into an existing account, the efficiency of the attack significantly increases. We present two approaches for choosing the attacker: a white-box gradient-based approach, and a black-box, model-free approach that relies on graph topology. Finally, we perform a comprehensive experimental evaluation of our approach on multiple datasets and GNN architectures.



(a) Before attacking: the target is classified as valid.

(b) After attacking: the target is classified as invalid.

Figure 1: An partial adversarial example from the test set of the Twitter dataset. An adversarially-crafted post perturbs the representation of the attacker node. This perturbation causes a misclassification of the target victim node, although they are not even direct neighbors.

2 Preliminaries

Let $\mathcal{G} = \{G_i\}_{i=1}^{N_G}$ be a set of graphs. Each graph $G = (\mathcal{V}, \mathcal{E}, \boldsymbol{X}) \in \mathcal{G}$ has a set of nodes $\mathcal{V}$ and a set of edges $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$, where $(u, v) \in \mathcal{E}$ denotes an edge from a node $u \in \mathcal{V}$ to a node $v \in \mathcal{V}$. $\boldsymbol{X} \in \mathbb{R}^{N \times D}$ is a matrix of D-dimensional node features. The i-th row of $\boldsymbol{X}$ is the feature vector of the node $v_i \in \mathcal{V}$ and is denoted as $\boldsymbol{x}_i = \boldsymbol{X}_{i,:} \in \mathbb{R}^D$.

Graph neural networks GNNs operate by iteratively propagating neural messages between neighboring nodes. Every GNN layer updates the representation of every node by aggregating its current representation with the current representations of its neighbors.

Formally, each node is associated with an initial representation $\boldsymbol{x}_v^{(0)} = \boldsymbol{h}_v^{(0)} \in \mathbb{R}^D$. This representation is considered as the given features of the node. Then, a GNN layer updates each node's representation given its neighbors, yielding $\boldsymbol{h}_v^{(1)} \in \mathbb{R}^{d_1}$ for every $v \in \mathcal{V}$. In general, the ℓ-th layer of a GNN is a function that updates a node's representation by combining it with its neighbors:

$$\boldsymbol{h}_v^{(\ell)} = \text{COMBINE}\left(\boldsymbol{h}_v^{(\ell-1)}, \{\boldsymbol{h}_u^{(\ell-1)} \mid u \in \mathcal{N}_v\}; \theta_\ell\right), \tag{1}$$

where $\mathcal{N}_v$ is the set of direct neighbors of v: $\mathcal{N}_v = \{u \in \mathcal{V} \mid (u, v) \in \mathcal{E}\}$.

The COMBINE function is what mostly distinguishes GNN types. For example, graph convolutional networks (GCN) (Kipf & Welling, 2017) define a layer as:

$$\boldsymbol{h}_v^{(\ell)} = \text{ReLU}\left(\sum\nolimits_{u \in \mathcal{N}_v \cup \{v\}} \frac{1}{c_{u,v}} \boldsymbol{W}^{(\ell)} \boldsymbol{h}_u^{(\ell-1)}\right) \tag{2}$$

where $c_{u,v}$ is a normalization factor usually set to $\sqrt{|\mathcal{N}_v| \cdot |\mathcal{N}_u|}$. After ℓ such aggregation iterations, every node representation captures information from all nodes within its ℓ-hop neighborhood. The total number of layers L is usually determined empirically as a hyperparameter. In the node classification scenario, we use the final representation $\boldsymbol{h}_v^L$ to classify v.

For brevity, we focus our definitions on the semi-supervised transductive node classification goal, where the dataset contains a single graph G, and the split into training and test sets is across nodes in the same graph. Nonetheless, these definitions can be trivially generalized to the inductive setting, where the dataset contains multiple graphs, the split into training and test sets is between graphs, and the test nodes are unseen during training.

We associate each node $v \in \mathcal{V}$ with a class $y_v \in \mathcal{Y} = \{1, ..., Y\}$. The labels of the training nodes are given during training; the test nodes are seen during training – without their labels. The training subset is represented as $\mathcal{D} = \left(G, \{(v_i, y_i)\}_{i=0}^{N_\mathcal{D}}\right)$. Given the training set, the goal is to learn a model $f_\theta : (\mathcal{G}, \mathcal{V}) \to \mathcal{Y}$ that will classify the rest of the nodes correctly. During training, the model f_θ thus minimizes the loss over the given labels, using $J(\cdot, \cdot)$, which typically is the cross-entropy loss:

$$\theta^* = \text{argmin}_\theta \mathcal{L}(f_\theta, \mathcal{D}) = \text{argmin}_\theta \frac{1}{N_\mathcal{D}} \sum\nolimits_{i=0}^{N_\mathcal{D}} J(f_\theta(G, v_i), y_i) \tag{3}$$

3 SINGLE-NODE GNN ATTACK

In this section, we describe our Single-node INdirect Gradient adversariaL Evasion (*SINGLE*) attack. While our attack is simple, it is the first attack that focuses on perturbing nodes (in contrast to edges (Dai et al., 2018)), which works with an *arbitrary single attacker node* (in contrast to multiple nodes (Zügner et al., 2018)) that is not the node under attack (in contrast to "direct" attacks where the attacker perturbs the node under attack directly (Zügner et al., 2018; Li et al., 2020)).

3.1 PROBLEM DEFINITION

Given a graph G, a trained model f_θ, a "victim" node v from the test set along with its classification by the model $\hat{y}_v = f_\theta(G, v)$, we assume that an adversary controls another node a in the graph. The goal of the adversary is to modify its own feature vector $\boldsymbol{x}_a$ by adding a perturbation vector $\boldsymbol{\eta} \in \mathbb{R}^D$ of its choice, such that the model's classification of v will *change*.

We denote by $G_{\boldsymbol{x}_a+\boldsymbol{\eta}}$ the graph G where the row of $\boldsymbol{X}$ that corresponds to the node a was added with the vector $\boldsymbol{\eta}$. In a non-targeted attack, the goal of the attacker is to find a perturbation vector $\boldsymbol{\eta}$ that will change the classification to *any* other class, i.e., $f_\theta(G_{\boldsymbol{x}_a+\boldsymbol{\eta}}, v) \neq f_\theta(G, v)$. In a *targeted* attack, the adversary chooses a specific label $y_{adv} \in \mathcal{Y}$ and the adversary's goal is to force $f_\theta(G_{\boldsymbol{x}_a+\boldsymbol{\eta}}, v) = y_{adv}$.

A modification of the features of a will affect the classification of v only if the *distance* in the graph between a and v is lower than or equal L – the number of GNN layers. Otherwise, a will not be contained in the receptive field of v, and the attack will result in "under-reaching" (Alon & Yahav, 2020) – any perturbation of a will not affect the prediction of v (Barceló et al., 2020). Therefore, we require that $\text{distance}_G(a, v) \leq L$.

In this work, we focus on gradient-based attacks. These kinds of attacks assume that the attacker can access a similar model to the model under attack and compute gradients. As recently shown by Wallace et al. (2020), this is reasonable assumption: an attacker can query the original model; using these queries, imitate the model under attack by training an imitation model; find adversarial examples using the imitation model; and transfer these adversarial examples back to the original model. Under this assumption, these attacks are general and are applicable to any GNN and dataset.

3.2 CHALLENGES

Unnoticeable Perturbations. Our first challenge is to find an adversarial example that will allow an imperceptible perturbation of the input. This objective is attainable in continuous domains such as images (Szegedy et al., 2013; Goodfellow et al., 2014) and audio (Carlini & Wagner, 2018) if we constrain l_∞-norm of the perturbation vector $\boldsymbol{\eta}$. It is, however, unclear what imperceptibility means in graphs. In most GNN datasets, a node's features are a bag-of-words representation of the words that are associated with the node. For example, in Cora (McCallum et al., 2000; Sen et al., 2008), every node is annotated by a many-hot feature vector of words that appear in the paper; in PubMed (Namata et al., 2012), node vectors are TF-IDF word frequencies; in Twitter (Ribeiro et al., 2017), node features are averages of GloVe embeddings, which can be viewed as word frequency vectors multiplied by a (frozen) embedding matrix. We argue that an attack would be unnoticeable in an academic paper or in a set of Tweets if the frequency of some words is slightly modified. For example, a particular word may be repeated a few times throughout the text or remain unused.

To constrain the $\boldsymbol{\eta}$ vector, we require that $\|\boldsymbol{\eta}\|_\infty \leq \epsilon_\infty$ – the maximal absolute value of the elements in the perturbation vector – is bounded by $\epsilon_\infty \in \mathbb{R}^+$.

Perturbing nodes instead of edges. Previous work mostly focused on perturbing graph *edges*. Zügner et al. (2018) perturb both edges and node features, but conclude that "perturbations in the structure lead to a stronger change in the surrogate loss compared to feature attacks"; Wu et al. (2019) also conclude that "perturbing edges is more effective than modifying the features". In this paper, we counter these conclusions and show that small node feature perturbations are stronger: (i) first, removing all the edges of a particular node is a special case of node feature perturbation. There exists a perturbation $\boldsymbol{\eta}$ such that $\boldsymbol{W}^1(\boldsymbol{x}_a + \boldsymbol{\eta}) = \boldsymbol{0}$, i.e., the modified feature vector $\boldsymbol{x}_a + \boldsymbol{\eta}$ is in the null space of the first GNN layer.[1] Such a feature perturbation is equivalent to *removing all the edges* of

[1]This equation demonstrates GCN, but similar equations hold for other GNN types like GAT and GIN.

the node a. (ii) Second, we argue that perturbing the graph structure is not realistic, because a single attacker controls only its own edges, and cannot control the *global* graph structure as in previous work (Dai et al., 2018; Bojchevski & Günnemann, 2019b; Zhang & Zitnik, 2020). (iii) Finally, when a successful attack is caused by removing edges, it is unclear whether the misclassification is caused by sensitivity to non-robust features in the data (Ilyas et al., 2019), or simply due to smaller amount of information. Similarly, when a successful attack is caused by inserting edges, it is unclear whether this is simply due to incorrect or unrealistic added information.

3.3 Finding the Perturbation Vector

To find the perturbation, we iteratively differentiate the desired loss of v with respect to the perturbation vector $\boldsymbol{\eta}$, update $\boldsymbol{\eta}$ according to the gradient, and add it to the feature vector. In non-targeted attacks, we take the positive gradient of the loss of the undesired label to increase the loss; in targeted attacks, we take the negative gradient of the loss of the adversarial label $\boldsymbol{y}_{adv}$:

$$\boldsymbol{\eta}^{t+1} = \begin{cases} \boldsymbol{\eta}^t + \gamma \nabla_{\boldsymbol{\eta}} J\left(f_\theta\left(G_{\boldsymbol{x}_a + \boldsymbol{\eta}^t}, v\right), \hat{y}_v\right) & \text{non-targeted attack} \\ \boldsymbol{\eta}^t - \gamma \nabla_{\boldsymbol{\eta}} J\left(f_\theta\left(G_{\boldsymbol{x}_a + \boldsymbol{\eta}^t}, v\right), y_{adv}\right) & \text{targeted attack} \end{cases} \tag{4}$$

where $\gamma \in \mathbb{R}^+$ is a learning rate. We repeat this process for a predefined number of K iterations, or until the model predicts the desired label.

Enforcing the constraints. We treat the node features as continuous throughout the attack iterations, whether they are discrete or continuous. Once the attack succeeds, we try to reset to zero as many perturbation vector elements as possible. We sort the perturbation vector elements in a decreasing order, according to their absolute value: $\{i_1, ..., i_D\}$. We start with the index of $\boldsymbol{\eta}$ whose absolute value is the largest, $\boldsymbol{\eta}_{i_1}$, and reset the rest of the $\{i_2, ..., i_D\}$ elements to zero. We then check whether perturbing only the i_1 index is sufficient. If the attack succeeds, we stop. If the attack fails (because of the large number of perturbation vector elements set to zero), we continue perturbing the rest of the elements of $\boldsymbol{\eta}$. In the worst case, we perturb all the D vector elements of $\boldsymbol{\eta}$. In most cases, we stop much earlier, practically perturbing only a small fraction of the vector elements.

Differentiate by frequencies, not by embeddings. When taking the gradient with respect to the perturbation vector $\nabla_{\boldsymbol{\eta}}$, there is a subtle, but crucial, difference between the way that node representations are given in the dataset: (a) *indicative* datasets provide initial node representations $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, ...]$ that are word indicator vectors (many-hot) or frequencies such as (weighted) bag-of-words (Sen et al., 2008; Shchur et al., 2018); (b) in *encoded* datasets, initial node representations are given encoded, e.g., as an average of word2vec vectors (Hamilton et al., 2017; Hu et al., 2020). *Indicative* datasets can be converted to *encoded* by multiplying every vector by an embedding matrix; *encoded* datasets *cannot* be converted to *indicative*, without the authors releasing the textual data that was used to create the *encoded* dataset.

In *indicative* datasets, a perturbation of a node vector *can* be realized as a perturbation of the original text from which the *indicative* vector was derived. That is, adding or removing words in the text can result in the perturbed node vector. In contrast, a few-indices perturbation in *encoded* datasets might be an effective attack, but will *not* be realistic because there is no perturbation of the original text that will result in that perturbation of the vector. That is, when perturbing nodes, it is crucial to use *indicative* datasets, or convert *encoded* datasets to the *indicative* representation from which they were derived (as we do in Section 4) using their original text.

4 Evaluation

We evaluate and analyze the effectiveness of our *SINGLE* attack. In Section 4.1, we show that *SINGLE* is more effective than alternatives such as single-edge attacks. In Section 4.2 we show that if we are allowed to *choose* the attacker node, *SINGLE* is much more effective.

Setup. Our implementation is based on PyTorch Geometric (Fey & Lenssen, 2019) and its provided datasets. We trained each GNN type with two layers ($L = 2$), using the Adam optimizer, early stopped according to the validation set, and applied a dropout of 0.5 between layers. In datasets where the node features are *discrete* (i.e., many-hot), we used the minimal possible perturbation of $\epsilon_\infty = 1$. Where the node features are *continuous* (e.g., frequencies of words), we used $\epsilon_\infty = 0.1$.

	Cora	CiteSeer	PubMed	Twitter
Clean (no attack)	80.5±0.8	68.5±0.7	78.5±0.6	92.4
EdgeGrad	70.5±1.3	52.6±1.0	63.0±1.4	83.5
SINGLE	**60.10**	**31.90**	**43.10**	**59.23**
SINGLE-hops	67.80	52.30	47.20	68.14

Table 1: Test accuracy under different types of attacks, when the attacker node is chosen *randomly*. Performed using GCN, and $\epsilon_\infty = 1$ for the discrete datasets (Cora and CiteSeer), and $\epsilon_\infty = 0.1$ for the continuous datasets (PubMed and Twitter).

An analysis of these values is presented in Section 4.4. We used up to $K = 20$ attack iterations. All experiments in this section were performed with GCN. Experiments with additional GNN types such as GAT and GraphSage are shown in Section 4.4.

Data. We used Cora and CiteSeer (Sen et al., 2008) which are *discrete* datasets, i.e., the given node feature vectors are many-hot vectors. Thus, we set $\epsilon_\infty = 1$, the minimal possible perturbation. We also used PubMed (Sen et al., 2008) and the Twitter-Hateful-Users (Ribeiro et al., 2017) datasets, which are *continuous*, and node features represent frequencies of words. Continuous datasets allow a much more subtle perturbation, and we set $\epsilon_\infty = 0.1$.

The Twitter-Hateful-Users dataset is originally provided as an *encoded* dataset, where every node is an average of GloVe vectors (Pennington et al., 2014). We reconstructed this dataset using the original text from Ribeiro et al. (2017), to be able to compute gradients with respect to the weighted histogram of words, rather than the embeddings. We took the most frequent 10,000 words as node features, and used GloVe-Twitter embeddings to multiply by the node features. Statistics of all dataset are provided in the supplementary material.

Baselines. In *SINGLE* (Section 3.3) the attacker node is selected randomly for each victim node, and the attack perturbs this node's features according to ϵ_∞. SINGLE-*hops* is a modification of *SINGLE* where samples the attacker node *only among nodes that are not neighbors*, i.e., the attacker and the victim are not directly connected ($(a, v) \notin \mathcal{E}$). We compare to additional approaches from the literature: *EdgeGrad* follows most previous work (Xu et al., 2019a; Li et al., 2020; Zügner & Günnemann, 2020): *EdgeGrad* randomly samples an attacker node as in *SINGLE*, and either inserts or removes a single edge from or to the attacker node, according to the gradient.[2] If both use a randomly selected attacker node, *EdgeGrad* is strictly stronger than the *GradArgmax* attack of Dai et al. (2018), which only *removes* edges. We run each approach 5 times with different random seeds for each dataset, and report the average and standard deviation. Due to time constrains, part of the results lack mean and standard deviations. We will include missing results in the revised version.

4.1 Main Results

Table 1 shows our main results for untargeted attacks across various datasets. As shown, *SINGLE* is more effective than *EdgeGrad* across all datasets. *SINGLE-hops*, which is more unnoticeable than attacking any arbitrary node, performs almost as good as *SINGLE* which uses an arbitrary node, and better than *EdgeGrad*. On Twitter, *SINGLE* reduces the test accuracy significantly better than *EdgeGrad*: 59.23% compared to 83.5%. Results for *targeted* attacks are shown in Appendix A.2.

As we explain in Section 3.3, *SINGLE* tries to find a perturbation vector in which the number of perturbed elements is minimal. We measured the number of vector elements that the attack had perturbed in practice. In PubMed, *SINGLE* used 76 vector elements on average, which are 15% of the elements in the feature vector. In Cora, *SINGLE* perturbed 717 elements on average, which are 50%. In CiteSeer, *SINGLE* used 1165 attributes on average, which are 31% of the features.

[2]This can be implemented easily using *edge weights*: training the GNN with weights of 1 for existing edges, adding all possible edges with weights of 0, and taking the gradient with respect to the vector of weights.

	Cora	CiteSeer	PubMed	Twitter
GlobalEdgeGrad	**15.0±9.1**	**12.8±1.6**	31.9±0.9	82.73
SINGLE+GradChoice	35.80	25.60	**4.10**	**6.43**
SINGLE+Topology	34.00	23.60	6.20	**6.43**

Table 2: Test accuracy when the adversary can *choose* the attacker node.

	Cora	CiteSeer	PubMed	Twitter
SINGLE-two attackers	62.60	9.23	44.80	61.47
SINGLE-direct	26.10	17.20	0.10	44.58
SINGLE	60.10	39.00	43.10	59.23

Table 3: Scenario ablation: test accuracy under different attacking scenarios.

4.2 Attacker Choice

If the attacker could *choose* its node, e.g., by hijacking an existing account in a social network, could they increase the effectiveness of the attack? We examine the effectiveness of two approaches for choosing the attacker node.

Gradient Attacker Choice (*GradChoice*) chooses the attacker node according to the largest gradient with respect to the node representations (for a non-targeted attack): $a^* = \text{argmax}_{a_i \in \mathcal{V}} \|\nabla_{\boldsymbol{x}_i} J\left(f_\theta\left(G, v\right), \hat{y}_v\right)\|_\infty$. The chosen attacker node is never the victim node itself.

Topological Attacker Choice (*Topology*) chooses the attacker node according to topological properties of the graph. As an example, we choose the neighbor of the victim node v with the smallest number of neighbors: $a^* = \text{argmin}_{a \in \mathcal{N}_v} |\mathcal{N}_a|$. The advantage of this approach is that the attacker choice is *model-free*: if the attacker cannot compute gradients, they can at least choose the most harmful attacker node, and then perform the perturbation itself using other non-gradient approaches such as in Waniek et al. (2018); Chang et al. (2020).

To perform a fair comparison, we compare these approaches with *GlobalEdgeGrad*, which is similar to *EdgeGrad* that can insert or remove an edge, with the difference that the chosen edge can be chosen *from the entire graph*.

Results. Results for these attacker choice approaches are shown in Table 2. The main results are that choosing the attacker node significantly increases the effectiveness of the *SINGLE* attack: for example, in Twitter, from 59.23% (Table 1) to 6.43% test accuracy (Table 2).

In datasets where the given initial node features are continuous (PubMed and Twitter), *SINGLE+GradChoice* and *SINGLE+Topology* are much more effective than *GlobalEdgeGrad*, showing the superiority of node perturbation over edge perturbation in the global view. *SINGLE+Topology* is almost as effective as *SINGLE+GradChoice* – both reduce test accuracy down to 6.43% on Twitter; *SINGLE+Topology* achieves 6.20% test accuracy on PubMed while *SINGLE+GradChoice* achieves 4.10%; although *SINGLE+Topology* is model-free.

Interestingly, *GradChoice* and *Topology* agree on the choice of attacker node for 83.9% of the nodes in Cora, 87.1% of the nodes in CiteSeer, 45% of the nodes in PubMed, and on 52% of the nodes in Twitter, showing that the node selection can sometimes be performed model-free.

In datasets where the initial node features are discrete (Cora and CiteSeer), i.e., many-hot vectors, *GlobalEdgeGrad* reduces the test accuracy *more* than *GradChoice* and *Topology*. We believe that the reason is the difficulty of two-step optimization in discrete datasets: for example, *GradChoice* needs to choose the node, and find the perturbation afterwards. Finding a perturbation for a discrete vector is more difficult than in continuous datasets, and the choice of the attacker node may not be optimal.

4.3 Scenario Ablation

The main scenario that we focus on in this paper is a *SINGLE* approach that always perturbs a *single* node, which is *not* the victim node ($a \neq v$). We now examine our *SINGLE* attack in other, easier,

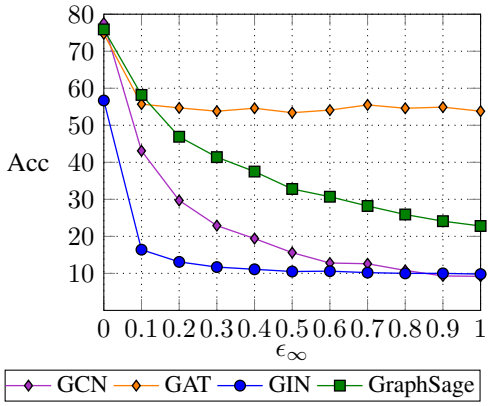


Figure 2: Effectiveness of the attack compared to the allowed ϵ_∞ (performed on PubMed, because its features are continuous).

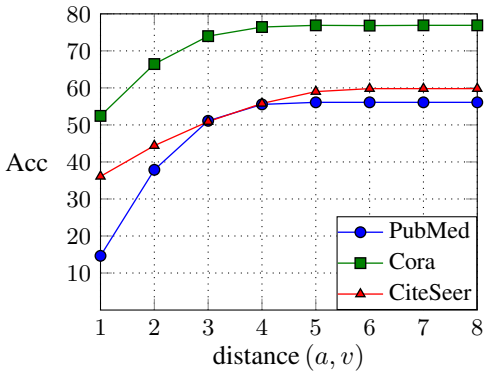


Figure 3: Test accuracy compared to the distance between the attacker node and the victim node, when the GCN was trained with $L = 8$ on PubMed.

scenarios: *SINGLE-two attackers* follows Zügner et al. (2018) and Zang et al. (2020), randomly samples *two* attacker nodes and perturbs their features using the same approach as *SINGLE*. *SINGLE-direct* perturbs the victim node directly (i.e., $a{=}v$), an approach that was found to be the most efficient by Zügner et al. (2018), albeit less realistic.

The results are shown in Table 3. Interestingly, perturbing *two* attacker nodes performs comparably to perturbing a single-node (*SINGLE*), across most datasets. This motivates our focus on the single-attacker scenario, and questions whether the multi-node attack assumptions of previous work are required. Expectedly, if we allow *SINGLE* to perturb the victim node directly (i.e., $a = v$), the attack is much more effective.

4.4 Sensitivity to ϵ_∞

How does the intensity of the adversarial perturbation affect the performance of the attack? Intuitively, we say that the less we restrict the perturbation (i.e., larger values of ϵ_∞), the more powerful the attack. We examine whether this holds in practice.

In our main experiments in Section 4, we used $\epsilon_\infty = 0.1$ for the continuous datasets (PubMed and Twitter). In this section, we vary the value of ϵ_∞ across different GNN types and observe the effectiveness of the attack. Figure 2 shows the results on PubMed. We used this dataset because it is larger than Cora and CiteSeer (Appendix A.1), and most importantly, its features are continuous, thus real-valued perturbations are feasible. As shown in Figure 2, the most significant difference is between performing the perturbation ($\epsilon_\infty = 0.1$) and not attacking at all ($\epsilon_\infty = 0$). As we increase the value of ϵ_∞, GCN and GraphSage (Hamilton et al., 2017) show a natural descent in test accuracy. Contrarily, GAT (Veličković et al., 2018) and GIN (Xu et al., 2019b) are more robust to increased absolute values of perturbations, while GAT is also the most robust compared to the other GNN types.

4.5 Distance Between Attacker and Victim

Does the effectiveness of the attack depend on the distance in the graph between the attacker and the victim? We trained a new model for each dataset using $L = 8$ layers. Then, for each test victim node, we sampled attackers according to their distance to the test node.

As shown in Figure 3, the effectiveness of the attack increases as the distance between the attacker and the victim decreases. At distance of 5, the curve seems to saturate and the test accuracy is close to its maximal value, i.e., the attack is ineffective across all datasets. A possible explanation for this is that apparently – more than few layers (e.g., $L = 2$ in Kipf & Welling (2017)) are not needed in most datasets. Thus, the rest of the layers can theoretically learn *not* to pass much of their input starting from the redundant layers, excluding adversarial signals as well.

5 RELATED WORK

Works on adversarial attacks on GNN differ in several main aspects. In this section, we discuss the main criteria, to clarify the settings that we address.

Single vs. multiple attackers All previous works allowed perturbing multiple nodes, or edges that are covered by multiple nodes: Zügner et al. (2018) perturb features of a *set* of attacker nodes; Zang et al. (2020) assume "a few bad actors"; other works perturb edges that in realistic settings their perturbation would require controlling multiple nodes (Bojchevski & Günnemann, 2019a; Sun et al., 2020; Chen et al., 2018). In this paper, we show that the limited scenario of a single independent attacker is as effective as multiple attacker nodes (Table 3).

Node vs. edge perturbations Most adversarial attacks on GNNs perturb the input graph by modifying the graph *structure* (Zügner & Günnemann, 2019; Wang et al., 2020; Xu et al., 2019a). For example, Dai et al. (2018) iteratively remove edges, yet their attack manages to reduce the accuracy by about 10% at most when perturbing a single edge. Li et al. (2020) also allow the insertion of edges; Waniek et al. (2018) and Chang et al. (2020) allow insertion and deletion of edges, using attacks that are based on correlations and eigenvalues, and not on gradients. Yefet et al. (2019) perturb one-hot node vectors, in the restricted domain of computer programs. Zügner et al. (2018) and Wu et al. (2019) perturb both edges and nodes; but they concluded that perturbing edges is more effective than perturbing nodes. In this work, we counter these conclusions and show that perturbing *node* features is more effective than perturbing edges.

Direct vs. influence attacks Another difference between prior works lies in the difference between *direct attacks* and *influence attacks*. In direct attacks, the attacker *perturbs the target node itself*. For example, the attack of Zügner et al. (2018) is the most effective when *the attacker and the target are the same node*. In influence attacks, the perturbed nodes are at least one hop away from the victim node. In this paper, we show that the strong *direct* assumption is not required (*SINGLE-direct* in Section 4.2), and that our attack is effective *when the attacker and the target are not even direct neighbors*, i.e., they are at least *two* hops away (*SINGLE-hops* in Section 4.1).

Poisoning vs. evasion attacks In a related scenario, some work (Zügner & Günnemann, 2019; Bojchevski & Günnemann, 2019a; Li et al., 2020; Zhang & Zitnik, 2020) focuses on *poisoning* attacks that perturb examples *before* training. Contrarily, we focus on the standard *evasion* scenario of adversarial examples in neural networks (Szegedy et al., 2013; Goodfellow et al., 2014), where the attack operates at test time, *after* the model was trained, as Dai et al. (2018).

Attacking vs. certifying Zügner & Günnemann (2020) focus on *certifying* the robustness of GNNs against adversarial perturbations; and Bojchevski & Günnemann (2019b) certified traditional PageRank-style models. In contrast, we study the effectiveness of the *adversarial attack* itself.

6 CONCLUSION

We demonstrate that GNNs are susceptible even to the extremely limited scenario of a single-node indirect adversarial example (*SINGLE*). The practical consequences of these findings are that a single attacker in a network can force a GNN to classify any *other* target node as the attacker's chosen label, by slightly perturbing some of the attacker's features. We further show that if the attacker can choose its attacker node – the effectiveness of the attack increases significantly. We study the effectiveness of these attacks across various GNN types and datasets.

We believe that this work will drive research in this field toward exploring novel defense approaches for GNNs. Such defenses can be crucial for real-world systems that are modeled using GNNs. Furthermore, we believe that the surprising results of this work motivate better theoretical understanding of the expressiveness and generalization of GNNs. To these ends, we make all our code and trained models publicly available.

REFERENCES

Uri Alon and Eran Yahav. On the bottleneck of graph neural networks and its practical implications. *arXiv preprint arXiv:2006.05205*, 2020.

Pablo Barceló, Egor V. Kostylev, Mikael Monet, Jorge Pérez, Juan Reutter, and Juan Pablo Silva. The logical expressiveness of graph neural networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=r1lZ7AEKvB.

Aleksandar Bojchevski and Stephan Günnemann. Adversarial attacks on node embeddings via graph poisoning. In *International Conference on Machine Learning*, pp. 695–704, 2019a.

Aleksandar Bojchevski and Stephan Günnemann. Certifiable robustness to graph perturbations. In *Advances in Neural Information Processing Systems*, pp. 8319–8330, 2019b.

Nicholas Carlini and David Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE Security and Privacy Workshops (SPW)*, pp. 1–7. IEEE, 2018.

Heng Chang, Yu Rong, Tingyang Xu, Wenbing Huang, Honglei Zhang, Peng Cui, Wenwu Zhu, and Junzhou Huang. A restricted black-box adversarial framework towards attacking graph embedding models. In *AAAI*, pp. 3389–3396, 2020.

Jinyin Chen, Yangyang Wu, Xuanheng Xu, Yixian Chen, Haibin Zheng, and Qi Xuan. Fast gradient attack on network embedding. *arXiv preprint arXiv:1809.02797*, 2018.

Hanjun Dai, Hui Li, Tian Tian, Xin Huang, Lin Wang, Jun Zhu, and Le Song. Adversarial attack on graph structured data. In *International Conference on Machine Learning*, pp. 1115–1124, 2018.

David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. Convolutional networks on graphs for learning molecular fingerprints. In *Advances in neural information processing systems*, pp. 2224–2232, 2015.

Matthias Fey and Jan E. Lenssen. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in neural information processing systems*, pp. 1024–1034, 2017.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. *arXiv preprint arXiv:2005.00687*, 2020.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems*, pp. 125–136, 2019.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017.

Jure Leskovec and Julian J Mcauley. Learning to discover social circles in ego networks. In *Advances in neural information processing systems*, pp. 539–547, 2012.

Jintang Li, Tau Xie, Liang Chen, Fentang Xie, Xiangnan He, and Zibin Zheng. Adversarial attack on large scale graph. *arXiv preprint arXiv:2009.03488*, 2020.

Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. Gated graph sequence neural networks. In *International Conference on Learning Representations*, 2016.

Andrew Kachites McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. Automating the construction of internet portals with machine learning. *Information Retrieval*, 3(2):127–163, 2000.

Alessio Micheli. Neural network for graphs: A contextual constructive approach. *IEEE Transactions on Neural Networks*, 20(3):498–511, 2009.

Galileo Namata, Ben London, L. Getoor, and Bert Huang. Query-driven active surveying for collective classification. 2012.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014. URL http://www.aclweb.org/anthology/D14-1162.

Manoel Horta Ribeiro, Pedro H Calais, Yuri A Santos, Virgílio AF Almeida, and Wagner Meira Jr. "Like sheep among wolves": Characterizing hateful users on twitter. *arXiv preprint arXiv:1801.00317*, 2017.

Manoel Horta Ribeiro, Pedro H Calais, Yuri A Santos, Virgílio AF Almeida, and Wagner Meira Jr. Characterizing and detecting hateful users on twitter. *arXiv preprint arXiv:1803.08977*, 2018.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2008.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *European Semantic Web Conference*, pp. 593–607. Springer, 2018.

Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Galligher, and Tina Eliassi-Rad. Collective classification in network data. *AI magazine*, 29(3):93–93, 2008.

Oleksandr Shchur, Maximilian Mumme, Aleksandar Bojchevski, and Stephan Günnemann. Pitfalls of graph neural network evaluation. *Relational Representation Learning Workshop, NeurIPS 2018*, 2018.

Yiwei Sun, Suhang Wang, Xianfeng Tang, Tsung-Yu Hsieh, and Vasant Honavar. Non-target-specific node injection attacks on graph neural networks: A hierarchical reinforcement learning approach. In *Proc. WWW*, volume 3, 2020.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Rakshit Trivedi, Hanjun Dai, Yichen Wang, and Le Song. Know-evolve: Deep temporal reasoning for dynamic knowledge graphs. In *International Conference on Machine Learning*, pp. 3462–3471, 2017.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJXMpikCZ.

Eric Wallace, Mitchell Stern, and Dawn Song. Imitation attacks and defenses for black-box machine translation systems. *arXiv preprint arXiv:2004.15015*, 2020.

Binghui Wang, Jinyuan Jia, Xiaoyu Cao, and Neil Zhenqiang Gong. Certified robustness of graph neural networks against adversarial structural perturbation. *arXiv preprint arXiv:2008.10715*, 2020.

Zhichun Wang, Qingsong Lv, Xiaohan Lan, and Yu Zhang. Cross-lingual knowledge graph alignment via graph convolutional networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 349–357, 2018.

Marcin Waniek, Tomasz P Michalak, Michael J Wooldridge, and Talal Rahwan. Hiding individuals and communities in a social network. *Nature Human Behaviour*, 2(2):139–147, 2018.

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 28th International Joint Conference on Artificial Intelligence*, pp. 4816–4823. AAAI Press, 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 28th International Joint Conference on Artificial Intelligence*, pp. 3961–3967. AAAI Press, 2019a.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019b. URL https://openreview.net/forum?id=ryGs6iA5Km.

Noam Yefet, Uri Alon, and Eran Yahav. Adversarial examples for models of code. *arXiv preprint arXiv:1910.07517*, 2019.

Xiao Zang, Yi Xie, Jie Chen, and Bo Yuan. Graph universal adversarial attacks: A few bad actors ruin graph learning models. *arXiv preprint arXiv:2002.04784*, 2020.

Xiang Zhang and Marinka Zitnik. Gnnguard: Defending graph neural networks against adversarial attacks. *arXiv preprint arXiv:2006.08149*, 2020.

Daniel Zügner and Stephan Günnemann. Certifiable robustness of graph convolutional networks under structure perturbations. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1656–1665, 2020.

Daniel Zügner and Stephan Günnemann. Adversarial attacks on graph neural networks via meta learning. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bylnx209YX.

Daniel Zügner, Amir Akbarnejad, and Stephan Günnemann. Adversarial attacks on neural networks for graph data. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2847–2856, 2018.

A SUPPLEMENTARY MATERIAL

Table 4: Dataset statistics.

	#Training nodes	#Val nodes	#Test nodes	#Unlabeled Nodes	#Classes
Cora	140	500	1000	2708	7
CiteSeer	120	500	1000	3327	6
PubMed	60	500	1000	19717	3
Twitter	4474	248	249	95415	2

A.1 DATASET STATISTICS

Statistics of the datasets that we used are shown in Table 4.

A.2 TARGETED ATTACKS

	Cora	CiteSeer	PubMed
EdgeGrad	29.50±1.3	47.40±1.0	37.00±1.4
SINGLE	**34.10**	**56.60**	**39.40**
SINGLE-indirect	29.30	45.80	36.00
GlobalGradEdge	**85.00**±9.1	87.20±1.6	68.10±0.9
SINGLE+GradChoice	80.40	98.90	49.10
SINGLE+Topology	57.40	**99.00**	**99.92**

Table 5: Attack success rate under different types of *targeted* attacks

Table 5 shows the success rate of *targeted* attacks across datasets and approaches. Differently from Table 1 which shows test accuracy, Table 5 presents the targeted attack's success rate (i.e., higher is better), which is the fraction of test examples that the attack managed to force *a specific label prediction*.