

Bandits for Structure Perturbation-based Black-box Attacks to Graph Neural Networks with Theoretical Guarantees

Binghui Wang^{*}, Youqi Li[†], and Pan Zhou[‡]

^{*}Department of Computer Science, Illinois Institute of Technology

[†]School of Cyberspace Science and Technology, and School of Computer Science, Beijing Institute of Technology

[‡]Hubei Engineering Research Center on Big Data Security, School of Cyber Science and Engineering, Huazhong University of Science and Technology

Email: ^{*}bwang70@iit.edu, [†]liyoubi@bit.edu.cn, [‡]panzhou@hust.edu.cn

Abstract

Graph neural networks (GNNs) have achieved state-of-the-art performance in many graph-based tasks such as node classification and graph classification. However, many recent works have demonstrated that an attacker can mislead GNN models by slightly perturbing the graph structure. Existing attacks to GNNs are either under the less practical threat model where the attacker is assumed to access the GNN model parameters, or under the practical black-box threat model but consider perturbing node features that are shown to be not enough effective. In this paper, we aim to bridge this gap and consider black-box attacks to GNNs with structure perturbation as well as with theoretical guarantees. We propose to address this challenge through bandit techniques. Specifically, we formulate our attack as an online optimization with bandit feedback. This original problem is essentially NP-hard due to the fact that perturbing the graph structure is a binary optimization problem. We then propose an online attack based on bandit optimization which is proven to be sublinear to the query number T , i.e., $\mathcal{O}(\sqrt{NT}^{3/4})$ where N is the number of nodes in the graph. Finally, we evaluate our proposed attack by conducting experiments over multiple datasets and GNN models. The experimental results on various citation graphs and image graphs show that our attack is both effective and efficient.

1. Introduction

Graph neural networks (GNNs) have been emerging as the most prominent methodology for learning with graphs, such as social networks, chemical networks, superpixel graphs, etc. GNNs have also advanced many graph-related applications including but not limited to drug discovery [24], fake news detection on social media [21], traffic forecasting [34], and superpixel graph classification [11]. However, recent works have shown that GNNs are vulnerable to ad-

versarial attacks [10, 37, 31, 38, 26, 32, 25, 20]. That is, an attacker can easily fool a GNN model by slightly perturbing the graph structure (e.g., injecting new fake edges into the graph or deleting the existing edges from the graph) or perturbing the node features. Most of the existing attacks to GNNs essentially rely on white-box or gray-box threat model [10, 37, 31, 38, 26, 32, 25]. An attacker can not only obtain the predictions generated by the targeted GNN model, but also know the whole (i.e., in white-box) or partial (i.e., in gray-box) GNNs' inner parameters and network structure. These threat models enable the attacker to derive the true gradients that can be used to construct an (almost) optimal edge/feature perturbation via first-order optimization approaches, e.g., projected gradient descent (PDG).

In practice, however, an attacker often has limited knowledge about the GNN model. For instance, many models are deployed as an API due to the commercial value. In these practical scenarios, an attacker can only obtain the model predictions by querying the API, while not knowing the model's other information. An attack based on such a realistic threat model is called a *black-box attack*, which significantly raises the bar for the attacker as he cannot obtain the gradient information. A recent work [20] performs black-box attacks against GNNs. However, this work has two key drawbacks. First, it assumes that the attacker can only perturb the (continuous) node features. Existing works (e.g., [37]) have shown that feature perturbation-based attacks to GNNs are significantly less effective than structure perturbation-based attacks. Second, the attack is implemented via a heuristic greedy algorithm, which has no theoretically guaranteed attack performance. Note that black-box attacks are classified as *soft-label* black-box attacks [14, 18] and *hard-label* [8] black-box attacks. The former means an attacker knows the confidence scores when querying a target model, while the latter means an attacker only knows the predicted label.

In this paper, we consider soft-label black-box attacks to GNNs with *structure perturbation*. However, such a new attack setting is much more challenging, as finding the optimal structure perturbation is essentially an NP-hard problem

[‡]Corresponding author.

(i.e., a binary optimization problem) and the attacker only obtains the predictions via querying the model. We take the first step to solve the structure perturbation-based black-box attacks to GNNs with theoretical guarantees. Specifically, we first reformulate our attack as a bandit optimization (i.e., online optimization with bandit feedback) problem, which characterizes the attacker’s query process on the black-box GNN model and captures the unknown gradients. Then, we handle the binary constraint of the discrete structure perturbation and integrate it into our bandit-based attack objective. Next, we design an efficient and effective online attack to GNNs. Finally, we theoretically analyze our attack. Our key contributions are summarized as follows:

- We design the first theoretically guaranteed structure perturbation-based black-box attacks to GNNs via bandit optimization.
- We prove that our bandit-based attack algorithm theoretically yields a sublinear regret bound $\mathcal{O}(\sqrt{NT}^{3/4})$ within T queries for attacking a graph with N nodes.
- We conduct extensive experiments to evaluate our attack over multiple graph datasets and GNN models and demonstrate the effectiveness and efficiency of our attack.

2. Preliminaries and Problem Formulation

2.1. Graph neural networks

Let $G = (\mathcal{V}, \mathcal{E}, \mathbf{X})$ be a graph, where $u \in \mathcal{V}$ is a node, $(u, v) \in \mathcal{E}$ is an edge between u and v , and $\mathbf{X} = [\mathbf{x}_1; \mathbf{x}_2; \dots; \mathbf{x}_N] \in \mathbb{R}^{N \times d}$ is the node feature matrix. We denote $\mathbf{a}_v = [a_{v1}; a_{v2}; \dots; a_{vN}] \in \{0, 1\}^N$ as the adjacency vector of node v . $N = |\mathcal{V}|$ and $M = |\mathcal{E}|$ are the number of nodes and edges, respectively. We denote d_u and $\mathcal{N}(u)$ as u ’s node degree and the neighborhood set of u . We consider GNNs for node classification in this paper¹. In this context, each node $u \in \mathcal{V}$ has a label y_u from a label set $\mathcal{Y} = \{1, 2, \dots, L_C\}$. Given a set of $\mathcal{V}_L \subset \mathcal{V}$ labeled nodes $\{(\mathbf{x}_u, y_u)\}_{u \in \mathcal{V}_L}$ as the training set, GNN for node classification is to learn a node classifier that maps each unlabeled node $u \in \mathcal{V} \setminus \mathcal{V}_L$ to a class $y \in \mathcal{Y}$ based on the graph G .

Generally speaking, GNN consists of two main steps: *neighborhood aggregation* and *node representation update*. Suppose a GNN has K layers. We denote v ’s representation in the k -th layer as $\mathbf{h}_v^{(k)}$, with $\mathbf{h}_v^{(0)} = \mathbf{x}_v$. In the neighborhood aggregation, GNN obtains the representation $\mathbf{l}_v^{(k)}$ by aggregating representations of v ’s neighbors in the $(k-1)$ -th layer as $\mathbf{l}_v^{(k)} = \text{AGG}(\{\mathbf{h}_u^{(k-1)} : u \in \mathcal{N}(v)\})$. In the node representation update, GNN updates v ’s representation at the k -th layer via combining v ’s previous layer’s representation $\mathbf{h}_v^{(k-1)}$ with the aggregated neighborhood’s representations $\mathbf{l}_v^{(k)}$: $\mathbf{h}_v^{(k)} = \text{UPDATE}(\mathbf{h}_v^{(k-1)}, \mathbf{l}_v^{(k)})$.

¹Our attack can be naturally generalized to GNNs for graph classification. We discuss This in Section 3.1.

Different GNNs use different AGG and UPDATE functions. For instance, in Graph Convolutional Network (GCN) [16], AGG is the element-wise mean pooling function and UPDATE is the ReLU activation function. More specifically, it has the following form:

$$\mathbf{h}_v^{(k)} = \text{ReLU}\left(\mathbf{W}^{(k)}\left(\sum_{u \in \mathcal{N}(v)} d_u^{-1/2} d_v^{-1/2} \mathbf{h}_u^{(k-1)}\right)\right),$$

where $\mathbf{W}^{(k)}$ is the parameters for the k -th layer. A node v ’s final representation $\mathbf{h}_v^{(K)} \in \mathbb{R}^{|\mathcal{Y}|}$ can capture the structural information of all nodes within v ’s K -hop neighbors. Moreover, the final node representations of training nodes are used to train the node classifier. Specifically, let $\Theta = \{\mathbf{W}^{(1)}, \dots, \mathbf{W}^{(K)}\}$ be the model parameters and v ’s output be $f_{\Theta}(\mathbf{a}_v) = \text{softmax}(\mathbf{h}_v^{(K)}) \in \mathbb{R}^{|\mathcal{Y}|}$, where $f_{\Theta}(\mathbf{a}_v)_y$ indicates the probability of node v being class y . Then, Θ are learnt by minimizing the cross-entropy loss on the outputs of the training nodes \mathcal{V}_L , i.e.,

$$\Theta^* = \arg \min_{\Theta} - \sum_{v \in \mathcal{V}_L} \ln f_{\Theta}(\mathbf{a}_v)_y. \quad (1)$$

With the learnt Θ^* , we can predict the label for each unlabeled nodes $u \in \mathcal{V} \setminus \mathcal{V}_L$ as $\hat{y}_u = \arg \max_y f_{\Theta^*}(\mathbf{a}_u)_y$.

2.2. Threat model

Attacker’s knowledge. The considered black-box attack setting in this paper implies that an attacker does not know the internal configurations (i.e., the learned parameters) of the targeted GNN model. For a target node $v \in \mathcal{V}$, the only information the attacker knows about the GNN model is the predictions $f_{\Theta^*}(\mathbf{a}_v)$ (i.e., output logits) via querying the GNN model f_{Θ^*} . Moreover, we also reasonably assume that the attacker knows her neighbors, i.e., the adjacency vector \mathbf{a}_v ³. In practice, the attacker naturally knows the neighbors of his controlled node. Taking social network as an instance, an attacker controls a malicious user, and this malicious user definitely knows his (non)neighbors in the social network. Note that the compared black-box RL-S2V attack [10] also requires that an attacker’s target node knows his neighbors.

Attacker’s capability. We consider that the attacker can modify the connection status (e.g., injecting new fake edges or removing the existing edges) between the target node v and other nodes in the graph. In practice, it also incurs different costs for the attacker to manipulate different edges. The attacker’s budget of manipulating edges is often limited, and we denote by C the cost budget. We also constrain that the number of edges to be manipulated is bounded by B .

Attacker’s goal. Based on the attacker’s knowledge and capability, an attacker’s goal is to fool a targeted GNN, i.e., making her target node v ’s predicted label different from the true label y_v , by perturbing her adjacency vector \mathbf{a}_v with the cost budget C and allowed number of perturbed edges B .

²Note that the prediction also depends on v ’s node feature \mathbf{x}_v and the whole graph G . We omit \mathbf{x}_v and G for notation simplicity.

³For graph classification, we assume attackers know the input graph.

Our threat model requires that an attacker knows the confidence scores (as many existing attacks to DNN models). Although it is stronger than the threat model that an attacker only needs to know the hard label, we also highlight that this is the first optimization-based attack that targets discrete graph structure perturbation, where this problem itself is rather challenging. We will leave addressing the attack with hard labels as the query feedback in future work.

2.3. Problem formulation

Given the target node v , label y_v , and adjacency vector \mathbf{a}_v , an attacker aims to modify the connection status related to the target node v such that the targeted GNN misclassifies v . Let $\mathbf{s}_v \in \{0, 1\}^N$ be the adversarial structure perturbation on v , where $s_{vu} = 1$ means the connection status between the nodes v and u is changed, and $s_{vu} = 0$, otherwise. Then, we define the perturbed adjacency vector for v as $\mathbf{a}_v \oplus \mathbf{s}_v$, where \oplus is the XOR operator between two binary vectors. Moreover, we denote $\mathbf{c}_v \in \mathbb{R}^N$ as the cost vector associated with v , i.e., c_{vu} is the cost to modify the connection status between v and u . In the focused black-box setting, we consider the untargeted attack. Let $L(\mathbf{a}_v)$ be the loss function for the targeted node v without attack. With the adversarial perturbation \mathbf{s}_v , we have the attack loss as $L(\mathbf{a}_v \oplus \mathbf{s}_v)$. In this paper, we use the CW attack loss function [3] with κ attack confidence. Specifically, it is defined as follows:

$$L(x) = \max\{f_{\Theta^*}(x)_y - \max_{\hat{y} \neq y} \{f_{\Theta^*}(x)_{\hat{y}}\}, -\kappa\}. \quad (2)$$

Finally, our problem of the structure perturbation-based black-box attack to GNN can be formulated as

$$\min_{\mathbf{s}_v} L(\mathbf{a}_v \oplus \mathbf{s}_v), \text{ s.t., } \mathbf{1}^T \mathbf{s}_v \leq B, \mathbf{c}^T \mathbf{s}_v \leq C, \mathbf{s}_v \in \{0, 1\}^N, \quad (3)$$

where the first constraint means the number of edges to be perturbed is no more than B and the second constraint means the total costs of the perturbation are no more than C .

Lemma 1. *Our problem in Eq. (3) is NP-hard.*

Proof. Our problem in Eq. (3) is a combinatorial optimization problem, actually a type of knapsack problem, which is a classical NP-hard problem. \square

Lemma 1 implies that it is difficult to calculate the optimal perturbation vector \mathbf{s}_v^* within polynomial time under large graphs (i.e., \mathbf{s}_v has large dimension). To this end, we aim to design an approximation algorithm to derive sub-optimal solution. One algorithm is to relax the combinatorial binary constraint $\mathbf{s}_v \in \{0, 1\}^N$ into convex hull $\mathbf{s}_v \in [0, 1]^N$ and obtain a continuous optimization problem. Let $\hat{\mathbf{s}}_v$ be the solution of the continuous optimization problem. We can derive the sub-optimal solution for the original problem in Eq. (3) by rounding $\hat{\mathbf{s}}_v$ into combinatorial space $\{0, 1\}^N$ using randomization sampling like Bernoulli sampling [32]. Then, we have the following lemma to characterize the relation between \mathbf{s}_v and $\hat{\mathbf{s}}_v$ in expectation:

Lemma 2. *When sampling \mathbf{s}_v element-wise in Bernoulli distribution using the probability from the relaxed vector $\hat{\mathbf{s}}_v \in [0, 1]^N$, then the expectation of \mathbf{s}_v is $\hat{\mathbf{s}}_v$, i.e., the condition $\mathbb{E}[\mathbf{s}_v] = \hat{\mathbf{s}}_v$ holds.*

Proof. This lemma holds due to the fact that a random variable \mathbf{X} subject to Bernoulli distribution on support $\{0, 1\}$ takes its probability as expectation, i.e., $\mathbb{E}[\mathbf{X}] = \mathbb{P}[\mathbf{X}]$. Applying this fact elements in \mathbf{s}_v , we can prove this lemma. \square

Conventionally, to solve our relaxed continuous optimization problem, we can apply the PGD approach by running gradient updates projected onto the feasible domain within several steps. However, PGD requires gradient information to be available. In our black-box attack setting, only the prediction result is available (by querying the GNN) instead of the exact gradient. Thus, the attack problem becomes how to estimate the gradient such that the PGD method can still be applied. It is shown that zeroth-order optimization (short for ZOO⁴) can be used to estimate the gradient [6, 15, 19]. However, ZOO suffers from a low convergence rate and high query overhead due to necessarily exploring all edges to estimate the gradient per round. We aim to estimate the unknown gradient by controlling the exploration-exploitation tradeoff via bandit methods. Reinforcement learning (RL) can also control the exploration-exploitation tradeoff. However, RL-based attack, i.e., RL-S2V [10], is naturally heuristic. Our attack can address both issues in ZOO-based and RL-based attacks. Specifically, our attack has theoretical guarantees and better attack performance than ZOO-based and RL-based attacks (See Section 6). Next, we reformulate our attack problem as a bandit optimization problem and then propose a solution to it.

2.4. Reformulating our attack as a bandit problem

When the attacker selects a perturbation vector \mathbf{s}_v and uses the perturbed adjacent vector $\mathbf{a}_v \oplus \mathbf{s}_v$ to query the GNN, the GNN returns the prediction $f_{\Theta^*}(\mathbf{a}_v \oplus \mathbf{s}_v)$. Thus, the objective $L(\mathbf{a}_v \oplus \mathbf{s}_v)$ is revealed based on Eq. (2), which can be seen as a bandit feedback (i.e., reward) for the selected perturbation vector \mathbf{s}_v (i.e., an arm). Under the bandit feedback, the attacker wants to maximize the cumulative rewards. Note that since the attacker does not know the optimal arm \mathbf{s}_v^* in each round, it will incur a regret, i.e., the difference of the maximum reward under the optimal arm \mathbf{s}_v^* in hindsight and the reward of the attacker's attack algorithm. Then, the attacker's goal is to minimize the cumulative regrets. Let $\text{Reg}(T)$ be the cumulative regrets in T rounds, and \mathbf{s}_v^t be the perturbation vector selected at round t , then the cumulative regrets $\text{Reg}(T)$ can be calculated as

⁴Note that ZOO is a perfect benchmark in our setting. First, only ZOO approximates the gradient directly through queries. Second, ZOO is the only method that also has a regret bound. Hence, we can compare with ZOO in terms of both theoretical results and empirical attack performance.

$\text{Reg}(T) = \mathbb{E}[\sum_{t=1}^T L(\mathbf{s}_v^t)] - TL(\mathbf{s}_v^*)$. In bandit optimization, it is important to design an arm selection algorithm with *sublinear* regret (i.e., $\text{Reg}(T) = o(T)$). This is because the selected arm \mathbf{s}_v at round T is asymptotically optimal when T is sufficiently large (i.e., $\lim_{T \rightarrow \infty} \frac{\text{Reg}(T)}{T} = 0$).

3. Structure Perturbation-based Black-Box Attacks to GNNs via Bandits

Here, we design an online attack to GNN based on bandits optimization and show its sublinear regret in next section.

First, we relax the binary perturbation vector $\mathbf{s}_v \in \{0, 1\}^N$ into a continuous convex hull $\hat{\mathbf{s}}_v \in [0, 1]^N$. In this case, we can define the arm set \mathcal{W} as: $\mathcal{W} = \{\hat{\mathbf{s}}_v \in [0, 1]^N \mid \mathbf{1}^T \hat{\mathbf{s}}_v \leq B, \mathbf{c}^T \hat{\mathbf{s}}_v \leq C\}$, where \mathcal{W} is convex. Note that our bandit for the black-box setting is different from the traditional multi-armed bandits (MAB), as arm set \mathcal{W} contains infinite perturbation vectors. Thus, the approaches like upper confidence bound (UCB) [2] and Thompson Sampling [22] in MAB fail to solve our problem. Moreover, it is impossible to use the combinatorial bandits [7, 29] to derive the optimal perturbation because they have to collect enough historical samples to calculate the mean of each perturbed edge. In particular, in the exploitation phase, they behave as ZOO [6] and estimate a gradient with $N + 1$ queries. In the exploration phase, they require an approximation algorithm to derive the suboptimal perturbation with the UCB values. However, to the best of knowledge, there is no such an efficient approximation algorithm in the context of structural perturbation attacks with theoretical guarantees.

Next, we need to address how to efficiently decide the next arm (i.e., $\hat{\mathbf{s}}_v^{t+1}$) at the end of round t such that the regret is minimized. We leverage online convex optimization (OCO) technique to derive the next arm $\hat{\mathbf{s}}_v^{t+1}$ at round $t + 1$. We note that the loss function $L(\cdot)$ is often non-convex. However, we emphasize that the gradient descent used in OCO is still useful as it is challenging to derive the closed-form solution for non-convex functions. OCO approach requires gradient information at the selected arm to conduct gradient descent. In our black-box attack setting, the attacker only receives the bandit feedback for the selected arm $\hat{\mathbf{s}}_v^t$. To estimate the gradient for the black-box attack at the selected arm $\hat{\mathbf{s}}_v^t$, the attacker can only query the GNN to compute the gradient. We use *one point gradient estimation* (OPGE) [13] technique to estimate the gradient due to its simplicity and effectiveness. The idea of OPGE is to find a vector in unit sphere $\mathcal{S} = \{\mathbf{u} \in \mathbb{R}^N \mid \|\mathbf{u}\|_2 = 1\}$ such that its direction has small intersection angle with the gradient (i.e., \mathbf{u} is a good approximation of the gradient). To this end, we uniformly sample a unit vector from \mathcal{S} and derive an approximate gradient in the following lemma.

Lemma 3. *For a unit vector \mathbf{u} uniformly sampled from the unit sphere \mathcal{S} and a sufficient small $\delta > 0$, we can estimate gradient as $\nabla \hat{L}(\hat{\mathbf{s}}_v) = \mathbb{E}_{\mathbf{u} \in \mathcal{S}}[\frac{N}{\delta} L(\hat{\mathbf{s}}_v + \delta \mathbf{u}) \mathbf{u}]$.*

Algorithm 1 Black-box attack to GNN for node classification via OCO with bandit

Input: Target GNN model f_{Θ}^* , target node v , maximal #perturbed edges B , cost budget C , PGD step η , $\delta > 0$, $\alpha \in [0, 1]$, query number T

Output: Perturbed vector \mathbf{s}_v

- 1: Initialize: $\mathbf{v}^1 = \mathbf{0} \in \mathcal{W} = \{\hat{\mathbf{s}}_v \in [0, 1]^N \mid \mathbf{1}^T \hat{\mathbf{s}}_v \leq B, \mathbf{c}^T \hat{\mathbf{s}}_v \leq C\}$
 - 2: **for** $t = 1$ **to** T **do**
 - 3: Attacker randomly chooses a unit vector \mathbf{u}^t from the unit sphere \mathcal{S} .
 - 4: Attacker determines a perturbation $\hat{\mathbf{s}}_v^t = \mathbf{v}^t + \delta \mathbf{u}^t$.
 - 5: Attacker converts $\hat{\mathbf{s}}_v^t$ to be binary \mathbf{s}_v^t by setting top- B values in $\hat{\mathbf{s}}_v^t$ to be 1 and others to be 0.
 - 6: Attacker queries the GNN model f_{Θ}^* with \mathbf{s}_v^t to obtain the predictions and CW loss $L(\mathbf{s}_v^t)$ using Eq. (2).
 - 7: Attacker conducts PGD and updates: $\mathbf{v}^{t+1} = \prod_{(1-\alpha)\mathcal{W}}(\mathbf{v}^t - \eta \hat{\mathbf{g}})$, $\hat{\mathbf{g}} = \frac{N}{\delta} L(\mathbf{s}_v^t) \mathbf{u}^t$.
 - 8: **end for**
 - 9: **return** \mathbf{s}_v^T
-

The details of our attack algorithm are shown in Algorithm 1. The inputs of our algorithm include the targeted GNN model f_{Θ}^* , target node v , maximal number of perturbed edges B , cost budget C , PGD step η , a small $\delta > 0$, projection scale $\alpha \in [0, 1]$, and query number T . The output is a perturbed vector \mathbf{s}_v for the target node v after T rounds. In line 1, we set $\mathbf{0}$ as the initial prior vector \mathbf{v}^1 . In line 2–8, we calculate a sub-optimal perturbed vector to attack the targeted GNN based on OCO with bandit feedback. At round t , we randomly select a unit vector \mathbf{u}^t from the unit sphere \mathcal{S} as a stochastic gradient in line 3. In line 4, we derive a relaxed perturbed vector $\hat{\mathbf{s}}_v^t$ by updating the prior vector \mathbf{v}^t according to the selected stochastic gradient \mathbf{u}^t . In line 5, we *convert $\hat{\mathbf{s}}_v^t$ to binary \mathbf{s}_v^t by setting its top- B nonzero values (which corresponds to the entries in $\hat{\mathbf{s}}_v^t$ with the B largest nonzero probabilities) to be 1 (thus perturbing at most B edges) and the remaining values to be 0*. In line 6, we query the GNN model f_{Θ}^* with \mathbf{s}_v^t and obtain a loss feedback $L(\mathbf{s}_v^t)$. In line 7, we conduct PGD on the arm set \mathcal{W} to update the \mathbf{v}^{t+1} for round $t + 1$. Finally, after T queries, we obtain the perturbed vector \mathbf{s}_v^T for the target node v .

3.1. Extending our attack for graph classification

Our proposed attack against node classification can be naturally extended to attack GNN models for graph classification with a small effort of modifications. In a graph classification model, its input is an adjacent matrix of a graph and its output is the label of the graph. In node classification, we aim to perturb the adjacent vector of a target node, while in graph classification, we perturb the adjacent matrix. Let $\mathbf{A} \in \{0, 1\}^{N \times N}$ be the adjacent matrix of a graph with N nodes. Moreover, let $\mathbf{S} \in \{0, 1\}^{N \times N}$ be a perturbation matrix, where $\mathbf{S}_{ij} = 1$ if the connection status between the edge (i, j) is modified, and $\mathbf{S}_{ij} = 0$, otherwise. To perform our attack against graph classification models, we only need to flatten the matrix \mathbf{S} into a vector \mathbf{s} and feed it as an input to our attack algorithm. After obtaining the perturbed \mathbf{s} , we can reshape it to a perturbed adjacent matrix.

4. Main Results

In this section, we analyze the regret bound of our attack algorithm where we assume the loss function is convex. We note that it is an interesting future work to generalize our analysis to non-convex loss functions. We first present the following assumptions and lemmas.

Assumption 1. *There exists a Lipschitz constant C_L such that the following inequality holds for any \mathbf{u} and \mathbf{v} ,*

$$|L(\mathbf{u}) - L(\mathbf{v})| \leq C_L \|\mathbf{u} - \mathbf{v}\|_2. \quad (4)$$

Lemma 4. *For continuous $\hat{\mathbf{s}}_v^t$ and the rounded binary \mathbf{s}_v , the instant regret is bounded as:*

$$\mathbb{E}[|L(\mathbf{s}_v) - L(\hat{\mathbf{s}}_v^t)|] \leq C_L N^{3/4} \sqrt{1 + \frac{\eta}{\delta}}. \quad (5)$$

Line 7 in Algorithm 1 can be seen as the stochastic gradient decent on $L(\hat{\mathbf{s}}_v + \delta \mathbf{u})$. Thus, we have the following lemma to characterize its regret bound.

Lemma 5 ([12]). *Suppose that the arm set \mathcal{W} satisfies $\mathcal{W} \subseteq R\mathbb{B}$ for given radius $R > 0$, where \mathbb{B} is unit ball, i.e., $\mathbb{B} = \{\mathbf{u} \in \mathbb{R}^N : \|\mathbf{u}\|_2 \leq 1\}$. When loss function $L(\cdot)$ is convex, the cumulative regret for relaxed continuous variables are bounded as follows:*

$$\mathbb{E}\left[\sum_{t=1}^T L(\hat{\mathbf{s}}_v^t)\right] - TL(\hat{\mathbf{s}}_v^*) \leq C_L R \sqrt{T}, \quad (6)$$

where $\hat{\mathbf{s}}_v^t$ is continuous variable at round t and $\hat{\mathbf{s}}_v^*$ is optimal continuous solution of our relaxed optimization problem.

Note that Lemma 5 just captures the regret on the whole arm set \mathcal{W} . However, line 7 in Algorithm 1 updates the arm $\hat{\mathbf{s}}_v^t$ by projecting onto set $(1 - \alpha)\mathcal{W}$ for $0 < \alpha < 1$. The incurred regret by conducting $(1 - \alpha)$ -projection is captured by the following lemma.

Lemma 6. *For time horizon T , the incurred regret due to $(1 - \alpha)$ -projection is bounded as follows:*

$$\min_{\mathbf{w} \in (1-\alpha)\mathcal{W}} \sum_{t=1}^T L(\mathbf{w}) - TL(\hat{\mathbf{s}}_v^*) \leq 2\alpha T. \quad (7)$$

Based on the above assumption and the lemmas, we have the following theorem on the regret bound of our attack:

Theorem 1. *Under T rounds attack span, our proposed attack algorithm incurs regret $\text{Reg}(T)$, which is upper bounded by $\mathcal{O}(\sqrt{NT}^{3/4})$ with T queries to the GNN.*

Remark. Theorem 1 not only demonstrates the sublinear regret our attack achieves, but also presents that our attack is dimension-efficient, i.e., $\mathcal{O}(\sqrt{N})$. It implies that our attack can be scalable to large graphs. Note that gradient-free ZOO [6] has query complexity $\mathcal{O}(N)$. From the regret bound, we can see that our attack enjoys a better convergence rate $\mathcal{O}(1/T^{3/4})$ than ZOO [19], which has a $\mathcal{O}(1/T^{1/2})$ convergence rate. Note also that Ilyas *et al.* [15] proposed a

bandit-based black-box attack to image classifiers, which can be adapted to solve our problem. However, their approach 1) does not provide theoretical results in terms of regret bound; and 2) is less efficient than ours due to requiring multiple gradient estimation in each iteration.

Computational complexity. Our attack requires 1 query per round and each query has time complexity $\mathcal{O}(N)$. ZOO requires $2N$ queries per round and each query has time complexity $\mathcal{O}(N)$ —Its time complexity per round is $\mathcal{O}(N^2)$. RL-S2V needs to train an extra Q -network, which is used to perform the attack. During the attack, it has the same time complexity as our attack: 1 query per round and each round has a time complexity $\mathcal{O}(N)$. These analyses show our attack is more efficient than RL-S2V and ZOO.

5. Related Work

Recent attacks to GNNs mainly focus on white-box/gray box settings [37, 10, 5, 38, 32, 27, 25, 4, 36, 20]. For instance, Zugner [37] proposed the first attack, called *Nettack*, against GCN for node classification by perturbing graph structure or/and node features. Specifically, *Nettack* learns a surrogate linear model of GCN by defining a graph structure-preserving perturbation that constrains the difference between the node degree distributions of the graph before and after an attack. Xu *et al.* [32] utilized the model gradient to generate perturbation on the topology of the graph. We study the black-box setting where gradient information is unknown to the attacker. A recent work [20] performed black-box attacks to GNNs. However, This work focus on perturbing the continuous node features, which does not fit our problem well and is less effective than discrete structure perturbation. In addition, the attack is implemented via a heuristic greedy algorithm, which has no theoretically guaranteed performance. Also, [37] has shown that feature perturbation-based attacks to GNNs is significantly less effective than structure perturbation-based attacks (e.g., an attacker needs to perturb an average of 100 node features, in order to have a comparable performance by the attack that perturbs only 5 edges). Thus, we focus on the graph structure perturbation in this paper. Zang *et al.* [35] studied the graph universal adversarial attacks where a set of anchor nodes is identified and their connection to the target node is flipped. However, how to select the optimal anchor nodes is not investigated. Some other works [28, 17] focus on attacking community detection and they are heuristic and orthogonal to our work.

Several black-box attacks [14, 15, 9] for non-graph classification models have been proposed. However, these methods cannot solve our problem because their attack problems are essentially continuous optimization problems. Note that [15] also used the bandit to formulate the black-box attack problem. However, their work does not have a theoretical regret bound and is less efficient due to using multiple gradient estimations in each iteration.

6. Experiments

6.1. Experimental Setup

Dataset description and GNN models. In node classification experiments, we use three benchmark citation graphs, i.e., Cora, Citeseer, and Pubmed [23]. In these graphs, each node represents a document and each edge indicates a citation between two documents. Each document treats the bag-of-words feature as the node feature vector, and has a label. We adopt the representative GCN [16] and SGC [30] for node classification, and evaluate our attack against the two models. In graph classification, we use two benchmark super-pixel graphs, i.e., MNIST and CIFAR10 [11], in computer vision. MNIST and CIFAR10 are classical image classification datasets. In our experiments, they are converted into graphs using the SLIC super-pixels [1]. Each node has the super-pixel coordinates as the feature and each super-pixel graph has a label. We adopt the representative GIN [33] for graph classification, and evaluate our attack against GIN. Table 1 summarizes the basic statistics of these datasets.

Training nodes/graphs and target nodes/graphs. We use the training nodes/graphs to train GNN models and use the target nodes/graphs to evaluate our attack against the trained models. Following existing works [16, 37, 11], in citation graphs, we randomly sample 20 nodes from each class as the training nodes; sample 100 nodes that are correctly classified by each GNN model as the target nodes. In image graphs, we respectively use 55,000 graphs and 45,000 graphs in MNIST and CIFAR10 for training, and randomly sample 100 graphs correctly classified by the GIN model as the target graphs.

Baselines. We compare our attack with two state-of-the-arts: *RL-based attack* [10] and *ZOO attack* [6]. Note that for the RL-based attack, we adjust the reward using the same CW attack loss like ours, and thus it has a fair comparison with our attack. We also choose *random attack* for comparison, where we generate structure perturbations by randomly changing connection status between pairs of nodes.

Cost simulation. We specify the cost of modifying the connection state for each pair of nodes. Note that the costs could be application-dependent. W.l.o.g., we assume the costs for different node pairs are uniformly distributed among a certain interval (e.g., [1,5] in our experiments). Note that the equal cost can be seen as a special case of the uniform cost.

Parameter setting. We set the hyperparameters in our attack algorithm as follows: $\eta = 10^{-4}$, $\delta = 10^{-6}$, and $\alpha = 0.7$ for attacking node classification methods, and $\eta = 10^{-1}$, $\delta = 10^{-3}$, and $\alpha = 0.6$ for attacking graph classification methods. Consider the different graph sizes, we set the default maximal number of perturbed edges B as 5 and 15 on the three citation graphs, and on the two image graphs, respectively; and the default total costs C is bounded by 25 on the citation graphs and 75 on the image graphs, respectively. In addition, the total number of queries $T = 50$ by

Table 1. Dataset statistics

Datasets	#Graphs	#Ave. Nodes	#Ave. Edges	#Classes	Task
Cora	1	2,708	5,429	7	(1)
Citeseer	1	3,327	4,732	6	(1)
Pubmed	1	19,717	44,338	3	(1)
MNIST	70K	70.57	564.53	10	(2)
CIFAR10	60K	117.63	941.07	10	(2)

(1) Node classification, (2) Graph classification.

default. We also study the impact of these parameters in our experiments. We implement our algorithm in Python and conduct experiments using public source codes. All our experiments are run in a computer with Intel(R) Core(TM) i7-6700 CPU @3.4Hz processors with 4 cores, 32GB RAM, 1 TB disk space and 6G GPU. We run all experiments 30 times and report the average result.

Evaluation metric. We adopt the attack successful rate, i.e., a fraction of the total (i.e., 100) targeted nodes/graphs misclassified after our attack, as the metric to measure the effectiveness of our attack. We also use the number of queries (T) to measure the efficiency of our attack. Specifically, we count each PGD iteration in our Algorithm 1 as a query.

6.2. Experimental Results

In this section, we provide a comprehensive evaluation of our black-box attack against GNNs for both node classification and graph classification. We aim to study our attack in terms of both effectiveness and efficiency. Our attack algorithm has three key factors: the number of perturbed edges B , total costs C , and the number of queries T . We will study the impact of these parameters one by one.

Impact of the number of perturbed edges. In this experiment, we separately fix the bounded total costs C to be 25 and 75, and set the number of queries T to be 50 on attacking both node classification models and graph classification models. The results of attacking these models are shown in Figure 1, Figure 2, and Figure 3, respectively.

We have the following observations: First, all attack approaches achieve a larger attack successful rate when the maximal number of perturbed edges B increases. This is because a larger B allows an adversary to have a better capability to perform the attack. Second, our attack significantly achieves a higher attack successful rate than the compared attacks in all GNNs models and graph datasets. For instance, when attacking GCN and SGC for node classification, our attack achieves the attack successful rate larger than 80% (and even 90%), while the second-best RL attack achieves the attack successful rate at most 80% across the three citation graphs. When attacking GIN for graph classification, our attack achieves 60% attack successful rates, while the RL attack obtains less than 53% attack successful rate in the two image graphs. All these results verify that the black-box bandit feedback in our algorithm is very useful to guide the selection of the optimal edges to be perturbed. In contrast, the heuristic RL attack is hard to do so.

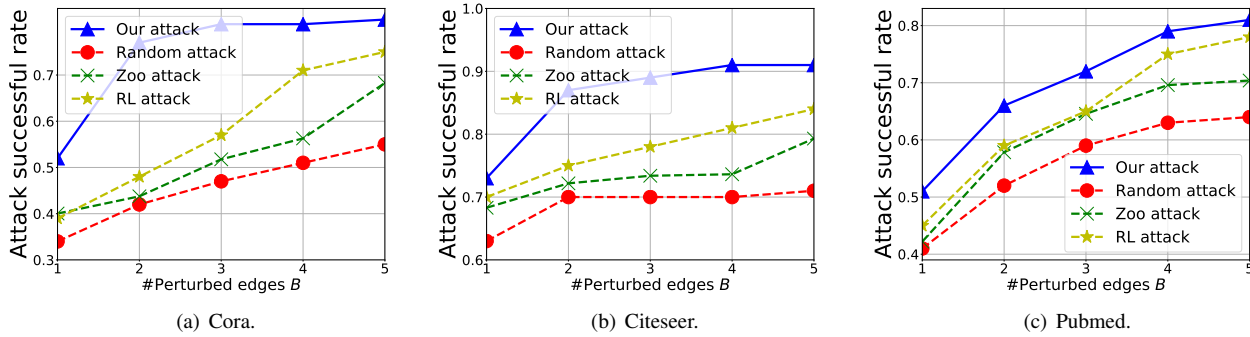


Figure 1. Attack successful rate vs. #perturbed edges B on GCN for node classification.

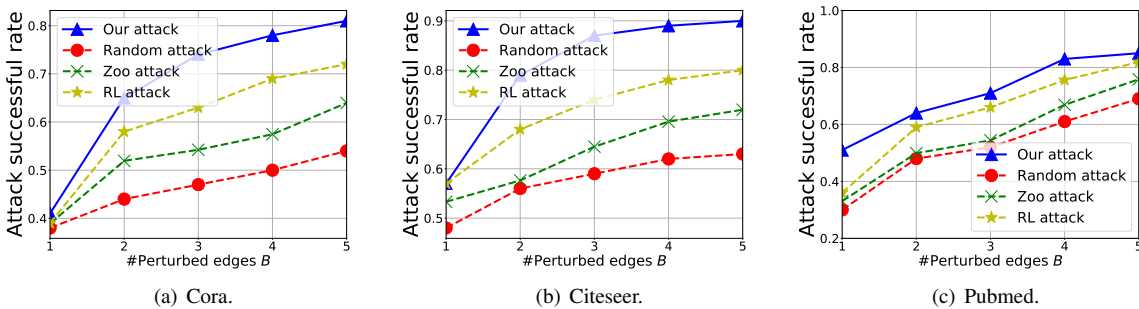


Figure 2. Attack successful rate vs. #perturbed edges B on SGC for node classification.

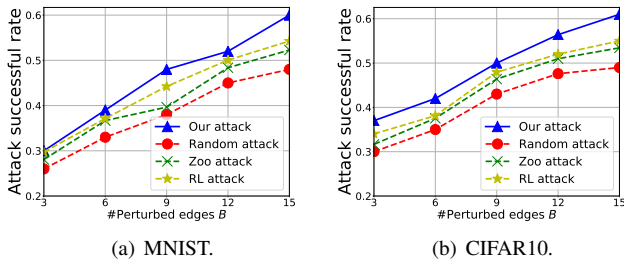


Figure 3. Attack successful rate vs. #perturbed edges B on GIN for graph classification.

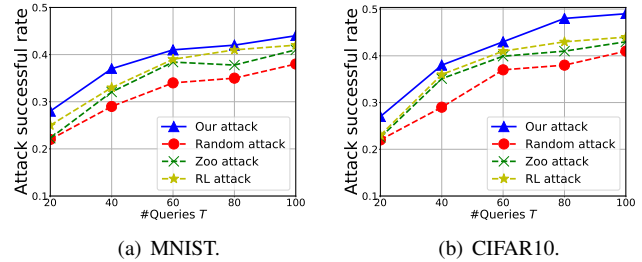


Figure 4. Attack successful rate vs. #queries T on GIN for graph classification.

Impact of the number of queries. In this experiment, we separately fix the bounded total costs C to be 25 and 75, and the maximal number of perturbed edges B to be 2 and 5 on attacking node classification models and graph classification models, respectively. The results of attacking GIN for graph classification and attacking GCN and SGC for node classification are shown in Figure 4, Figure 5, and Figure 6, respectively. First, all attack approaches achieve a larger attack successful rate when the number of queries T increases. This is because a larger T allows an adversary to obtain more predictions via querying the GNN models and thus have a better capability to perform the attack. Second, our attack requires much fewer queries to achieve the same attack successful rate than the compared attacks. In particular,

compared to our attack, RL attack requires 1.5-3x queries, when achieving comparable attack performance against node classification models and graph classification models, respectively. Third, given the same number of queries, our attack achieves 10% – 20% higher successful rate than RL attack across all citation graphs and image graphs. Similarly, these results again verify that the black-box bandit feedback is really beneficial in designing our attack.

Impact of the total costs. We study the impact of the total costs C and fix the number of queries T to be 50. As our attack outperforms the compared attacks, we only study our attack on attacking GCN for simplicity. Note that we have similar observations on attacking SGC and GIN.

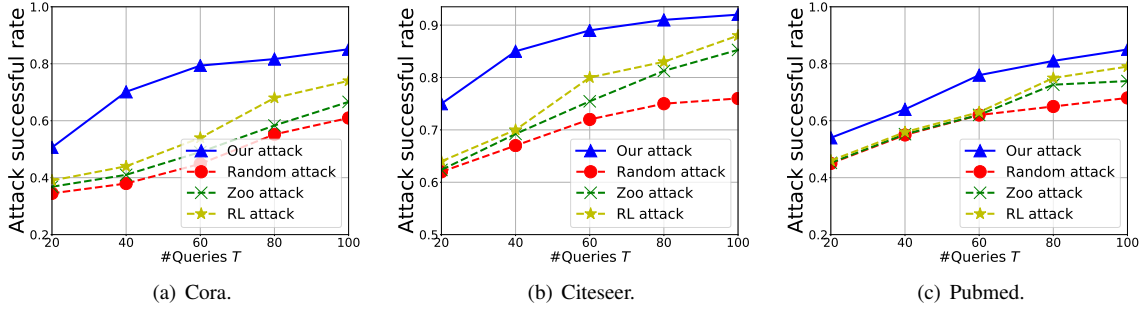


Figure 5. Attack successful rate vs. #queries T on GCN for node classification.

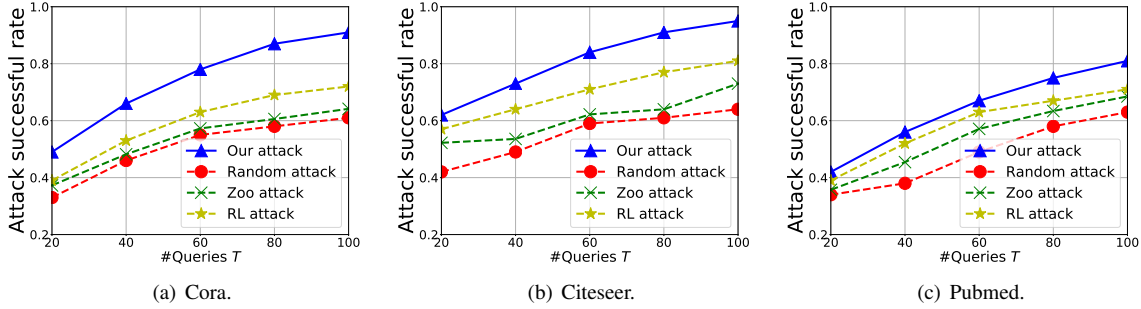


Figure 6. Attack successful rate vs. #queries T on SGC for node classification.

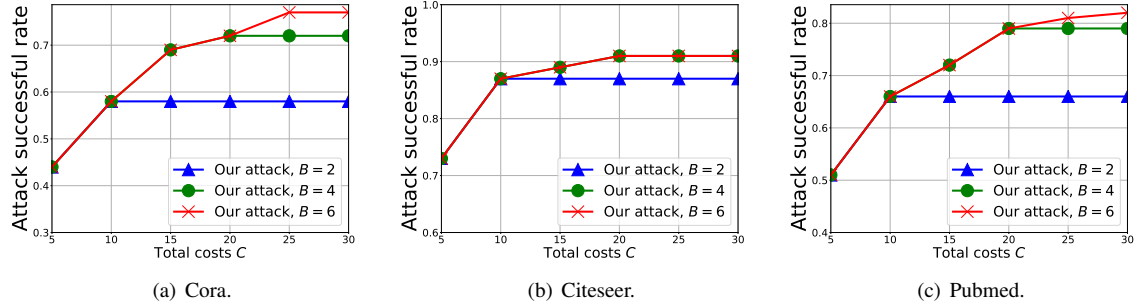


Figure 7. Attack successful rate vs. #total costs C on GCN for node classification.

The results are shown in Figure 7. We observe that the attack successful rate becomes higher when the cost budget is larger. Moreover, note that the cost for perturbing each edge is within $[1, 5]$. If the cost budget is sufficient, i.e., $C \geq 10, 20, 30$ for $B = 2, 4, 6$, then the attack successful rate is stable, as our attack finds the same space for structure perturbation. In contrast, if the cost budget is insufficient, even the allowed number of perturbed edges is larger, our attack performance cannot be improved, e.g., $C = 20$ for $B = 4, 6$ on Citeseer. The result demonstrates that cost budget is a key factor to affect our attack performance.

7. Conclusion and Future work

In this paper, we study black-box attacks to GNNs via manipulating the graph structure. We first formulate our

attack as a binary optimization problem, which is NP-hard. Then, we relax and reformulate our attack problem as a bandit optimization problem, and propose a bandit-based attack algorithm and rigorously prove that our attack yields a sublinear regret bound $\mathcal{O}(\sqrt{NT}^{3/4})$ within T queries for attacking a graph with N nodes. Finally, we evaluate our attack against GNN models for both node classification and graph classification. Our results demonstrate both the effectiveness and efficiency of our attack and that our attack significantly outperforms the state-of-the-arts.

Acknowledgments. This work of Wang is supported by the startup funding. The work of Li is partially supported by the National Natural Science Foundation of China (NSFC) under Grant No. 62102028, and China Postdoctoral Science Foundation under Grant No. 2021M700434, and Zhou is supported by the NSFC under Grant No. 61972448.

References

- [1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, and et al. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE TPAMI*, 2012.
- [2] Peter Auer and Ronald Ortner. Ucb revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61(1-2):55–65, 2010.
- [3] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE S&P*, 2017.
- [4] 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*, 2020.
- [5] Jinyin Chen, Yangyang Wu, Xuanheng Xu, Yixian Chen, Haibin Zheng, and Qi Xuan. Fast gradient attack on network embedding. *arXiv*, 2018.
- [6] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *AISec*, 2017.
- [7] Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit: General framework and applications. In *ICML*, 2013.
- [8] Minhao Cheng, Thong Le, Pin-Yu Chen, Jinfeng Yi, Huan Zhang, and Cho-Jui Hsieh. Query-efficient hard-label black-box attack: An optimization-based approach. In *ICLR*, 2019.
- [9] Shuyu Cheng, Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Improving black-box adversarial attacks with a transfer-based prior. In *NeurIPS*, 2019.
- [10] Hanjun Dai, Hui Li, Tian Tian, Xin Huang, Lin Wang, Jun Zhu, and Le Song. Adversarial attack on graph structured data. In *ICML*, 2018.
- [11] Vijay Dwivedi, Chaitanya K Joshi, Thomas Laurent, and et al. Benchmarking graph neural networks. *arXiv*, 2020.
- [12] Abraham D Flaxman, Adam Tauman Kalai, and H Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *SODA*, 2005.
- [13] ON Granichin. Stochastic approximation with input perturbation under dependent observation noises. *Vestn. Leningrad. Gos Univ*, (4):27–31, 1989.
- [14] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *ICML*, 2018.
- [15] Andrew Ilyas, Logan Engstrom, and Aleksander Madry. Prior convictions: Black-box adversarial attacks with bandits and priors. In *ICLR*, 2019.
- [16] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017.
- [17] Jia Li, Honglei Zhang, Zhichao Han, Yu Rong, Hong Cheng, and Junzhou Huang. Adversarial attack on community detection by hiding individuals. In *The Web Conference*, 2020.
- [18] Yandong Li, Lijun Li, and Liqiang Wang. Nattack: Learning the distributions of adversarial examples for an improved black-box attack on deep neural networks. In *ICML*, 2019.
- [19] Sijia Liu, Xingguo Li, Pin-Yu Chen, Jarvis Haupt, and Lisa Amini. Zeroth-order stochastic projected gradient descent for nonconvex optimization. In *IEEE GlobalSIP*, 2018.
- [20] Jiaqi Ma, Shuangrui Ding, and Qiaozhu Mei. Towards more practical adversarial attacks on graph neural networks. *NeurIPS*, 2020.
- [21] Federico Monti, Fabrizio Frasca, Davide Eynard, Damon Mannon, and Michael M Bronstein. Fake news detection on social media using geometric deep learning. *arXiv*, 2019.
- [22] Daniel Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, and Zheng Wen. A tutorial on thompson sampling. *Foundations and Trends in Machine Learning*, 2018.
- [23] Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Galligher, and Tina Eliassi-Rad. Collective classification in network data. *AI magazine*, 2008.
- [24] Chence Shi, Minkai Xu, Zhaocheng Zhu, and et al. Graphaf: a flow-based autoregressive model for molecular graph generation. In *ICLR*, 2020.
- [25] Yiwei Sun, Suhang Wang, Xianfeng Tang, Tsung-Yu Hsieh, and Vasant Honavar. Adversarial attacks on graph neural networks via node injections: A hierarchical reinforcement learning approach. In *The Web Conference*, 2020.
- [26] Binghui Wang and Neil Zhenqiang Gong. Attacking graph-based classification via manipulating the graph structure. In *CCS*, 2019.
- [27] Binghui Wang, Tianxiang Zhou, Minhua Lin, Pan Zhou, Ang Li, Meng Pang, Hai Li, and Yiran Chen. Efficient evasion attacks to graph neural networks via influence function. *arXiv*, 2020.
- [28] Marcin Waniek, Tomasz P Michalak, Michael J Wooldridge, and Talal Rahwan. Hiding individuals and communities in a social network. *Nature Human Behaviour*, 2018.
- [29] Zheng Wen, Branislav Kveton, Michal Valko, and Sharan Vaswani. Online influence maximization under independent cascade model with semi-bandit feedback. In *NIPS*, 2017.
- [30] Felix Wu, Tianyi Zhang, Amauri Holanda de Souza Jr, Christopher Fifty, Tao Yu, and Kilian Q Weinberger. Simplifying graph convolutional networks. In *ICML*, 2019.
- [31] Huijun Wu, Chen Wang, Yuriy Tyshetskiy, Andrew Docherty, Kai Lu, and Liming Zhu. Adversarial examples on graph data: Deep insights into attack and defense. In *IJCAI*, 2019.
- [32] Kaidi Xu, Hongge Chen, Sijia Liu, and et al. Topology attack and defense for graph neural networks: An optimization perspective. In *IJCAI*, 2019.
- [33] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *ICLR*, 2019.
- [34] Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In *IJCAI*, 2018.
- [35] Xiao Zang, Yi Xie, Jie Chen, and Bo Yuan. Graph universal adversarial attacks: A few bad actors ruin graph learning models. *arXiv*, 2020.
- [36] Zaixi Zhang, Jinyuan Jia, Binghui Wang, and Neil Zhenqiang Gong. Backdoor attacks to graph neural networks. In *SACMAT*, 2021.
- [37] Daniel Zügner, Amir Akbarnejad, and Stephan Günnemann. Adversarial attacks on neural networks for graph data. In *KDD*, 2018.
- [38] Daniel Zügner and Stephan Günnemann. Adversarial attacks on graph neural networks via meta learning. *ICLR*, 2019.