A General Black-box Adversarial Attack on Graph-based Fake News Detectors

Peican Zhu[1], **Zechen Pan**[2], **Yang Liu**[1*], **Jiwei Tian**[3], **Keke Tang**[4*] and **Zhen Wang**[1*]
[1]School of Artificial Intelligence, Optics and Electronics, Northwestern Polytechnical University
[2]School of Computer Science, Northwestern Polytechnical University
[3]Air Traffic Control and Navigation College, Air Force Engineering University
[4]Cyberspace Institute of Advanced Technology, Guangzhou University
{ericcan, w-zhen}@nwpu.edu.cn, 928598047@mail.nwpu.edu.cn, {yangliuyh, tangbohutbh}@gmail.com, tianjiwei2016@163.com

Abstract

Graph Neural Network (GNN)-based fake news detectors apply various methods to construct graphs, aiming to learn distinctive news embeddings for classification. Since the construction details are unknown for attackers in a black-box scenario, it is unrealistic to conduct the classical adversarial attacks that require a specific adjacency matrix. In this paper, we propose the first general black-box adversarial attack framework, i.e., General Attack via Fake Social Interaction (GAFSI), against detectors based on different graph structures. Specifically, as sharing is an important social interaction for GNN-based fake news detectors to construct the graph, we simulate sharing behaviors to fool the detectors. Firstly, we propose a fraudster selection module to select engaged users leveraging local and global information. In addition, a post injection module guides the selected users to create shared relations by sending posts. The sharing records will be added to the social context, leading to a general attack against different detectors. Experimental results on empirical datasets demonstrate the effectiveness of GAFSI.

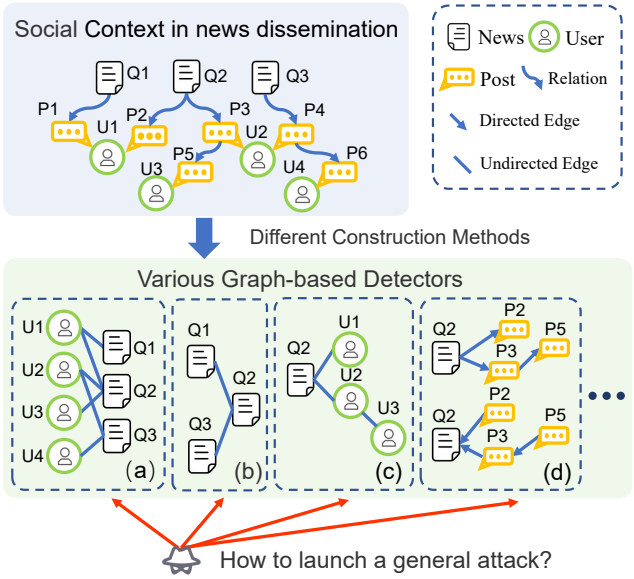


Figure 1: Social context in news dissemination can be constructed into (a) the user-news bipartite graph; (b) the news engagement graph; (c) the news-user propagation tree; and (d) the news-post propagation and dispersion tree, etc. The diversity in the graph types poses challenges for launching a general black-box adversarial attack on fake news detectors based on different graphs.

1 Introduction

The rapid expansion of online social media platforms has led to a rise in misinformation, undermining public trust in truth and science. Unlike traditional media, where content is often rigorously fact-checked, the interactive nature of social media accelerates the spread of fake news through commenting and sharing, magnifying its impact. This makes detecting and countering fake news on these platforms both challenging and crucial [Aïmeur *et al.*, 2023; Cheng *et al.*, 2024].

To tackle this challenge, abundant machine-learning methods are proposed for fake news detection. Besides utilizing single-modal or multi-modal detectors to extract features from content [Dong *et al.*, 2023; Hua *et al.*, 2023; Dong *et al.*, 2024], there is a growing interest in conceptualizing the online social network as a graph structure to leverage the rich social context [Phan *et al.*, 2023; Yin *et al.*, 2024]. Among existing studies, the user-news bipartite graph is commonly used for detectors to model user engagement [Nguyen *et al.*, 2020; Wang *et al.*, 2023a; Su *et al.*, 2023]. Besides, the news engagement graph can be constructed to explore relations between news directly [Wu and Hooi, 2023]. Instead of capturing information in the global network, the news's local propagation structure is also investigated. The news-post propagation and dispersion tree [Bian *et al.*, 2020] are constructed to aggregate information from two directions, while the news-user propagation tree [Dou *et al.*, 2021] is constructed to capture user-aware in the propagation structure. Due to the diversity of the graph construction, Graph Neural Network (GNN) is allowed to learn distinctive news embeddings from different perspectives for news classification. Despite their effectiveness in detecting fake news, these detectors have increasingly been found to be

*Corresponding authors.

vulnerable to adversarial attacks [Wang *et al.*, 2023a]. This vulnerability may be utilized to manipulate public opinion or gain financial benefits. Therefore, it is critical to investigate adversarial attacks on GNN-based fake news detectors to assess and enhance their robustness.

Existing adversarial attacks on GNN are primarily categorized into two types: edge perturbation and node injection. Edge perturbation techniques aim to alter the model's prediction by adding or removing edges in the graph [Zügner *et al.*, 2018; Li *et al.*, 2021; Zhang *et al.*, 2023], while node injection approaches focus on generating and introducing malicious nodes into the graph [Tao *et al.*, 2021; Wang *et al.*, 2022]. Although these methods show promise when the target graph is exposed, their efficacy diminishes in a more realistic black-box setting where the graph construction process is unknown. Given the vast differences in fake news detectors based on various social network structures, assuming a certain graph to conduct attacks becomes impractical for attackers. This situation leads to a crucial question: Is it possible to develop a universal attack strategy that remains effective in a black-box scenario against fake news detectors based on diverse social network structures?

In this paper, we propose a General Attack via Fake Social Interaction (GAFSI) against GNN-based fake news detectors, aiming to alter the prediction of the target news. Unlike existing studies, that are concentrated on a certain graph, we attempt to add social interaction records into the social context and thus can influence the graph representation regardless of the construction methods. Specifically, we employ a fraudster selection module to identify influential fraudsters for creating fraudster-post pairs. Then, through a post injection module, we select the source of new posts in the propagation structure and clone content from existing posts to new ones. Accordingly, the selected fraudsters will send the posts to simulate social interactions. This integration of social interaction records can effectively perturb the social context and fool GNN-based fake news detectors. We validate our framework's effectiveness in attacking various GNN-based fake news detectors with different graph structures. Our extensive experimental results demonstrate that GAFSI is general and capable of deceiving different detectors in a black-box setting, outperforming state-of-the-art methods.

Overall, our contribution is summarized as follows:

- We propose a general black-box attack framework to simulate social interactions, enhancing the effectiveness of the perturbation in a more realistic scenario.
- We propose GAFSI which leveraging gradient-based attention information, consists of a fraudster selection module and a post injection module.
- Extensive experimental results demonstrate that the perturbation generated by GAFSI on social context can effectively change the prediction of the target news.

2 Related Work

2.1 GNN-based Fake News Detection

According to the graph prototypes, GNN-based fake news detectors are primarily categorized into two types: propagation-based detectors and social-context-based detectors [Phan *et al.*, 2023]. The propagation-based detectors employ a tree-structured graph to model the propagation process of a single news [Bian *et al.*, 2020; Dou *et al.*, 2021; Silva *et al.*, 2021; Han *et al.*, 2021]. Here, the root node represents a news article and other nodes represent associated posts, while edges depict shared relations among nodes. Besides, social-context-based detectors prefer to capture global information. Recent work [Nguyen *et al.*, 2020] considers the relations among publishers, news, and users to construct the social context graph. Su *et al.* [2023] construct a dual-layer graph that contains a news propagation layer and a user interaction layer. They use shared relations to connect the two layers. Wu and Hooi [2023] construct a news engagement graph based on the number of common users. Although some methods utilize additional information to construct the graph, it is worth noting that shared relation remains a key factor in those graph construction processes. Hence, it inspired us to fool the GNN-based fake news detector via fake social interaction.

2.2 Adversarial Attack on GNN

Numerous adversarial attack methods have been proposed to fool classifiers in various domains [Zhu *et al.*, 2023; Tang *et al.*, 2023; Zhu *et al.*, 2024; Guo *et al.*, 2024]. For graph-related tasks, the attack techniques include attribute perturbation, edge perturbation, and node injection [Liu *et al.*, 2022]. Zügner *et al.* [2018] first conduct targeted attacks on GNN through greedy-based edge perturbation. To address the inefficiency of the previous methods on large-scale graphs, Li *et al.* [2021] introduce a method containing a subgraph constructing process. Zhang *et al.* [2023] focus on finding a minimized set of perturbations to realize an attack with low cost. Recent work [Shang *et al.*, 2023] start to conduct edge perturbation on heterogeneous graphs. They consider that all GNNs take the same graph as input and focus on perturbing the edge that has a larger common weight. Instead of edge perturbation, Tao *et al.* [2021] propose a targeted attack method by injecting a single node which is obtained from a pre-trained encoder. Wang *et al.* [2022] regard the node injection attack as a graph clustering problem and solved it based on Euclid's distance between nodes' adversarial feature. Although existing attacks perform well in traditional graph tasks, they demonstrate limitations when attacking GNN-based fake news detectors. The edge perturbation method may break the chain, leading to the destruction of the propagation structure, which is easy to discover. Furthermore, the node injection method generates perturbation in the form of embedding which can not be restored to original attributes. Hence, they are insufficient to conduct attacks against fake news detectors.

Wang *et al.* [2023a] is the first to attack GNN-based fake news detectors, focusing on manipulating three types of fraudsters to collaborate in deceiving the detector. They formulate the attack problem as attacking GNNs on a user-news bipartite graph and successfully alter the predicted label of the target news. However, they only consider modifying the edge between fraudsters and target news in the user-news bipartite graph. Thus, similar to previous attacks, their method can not attack detectors based on other graph structures.

3 Preliminary and Problem Statement

Social Context in News Dissemination. The social context in news dissemination is represented as $G = (U, Q, P, E)$, where $U = (u_1, \ldots, u_m)$ indicates a set of m users, $Q = (q_1, \ldots, q_n)$ denotes a set of n news, $P = (p_1, \ldots, p_l)$ represents a set of l posts, $E = \{E_u, E_t\}$ represents a set of relation in G. Here, E_u characterizes the user-post relations, with an edge $(u_i, p_j) \in E_u$ signifying that user u_i is the author of post p_j. E_t characterizes a set of the shared relations between news and posts, as well as the shared relations among posts themselves. Then, the fake news detection task based on G can be treated as a binary classification problem. $y_i \in Y$ is the ground-truth label of the news i where 0 and 1 represent the real news and fake ones respectively.

Graph Construction of G for Fake News Detection. To mine rich information in social context G, GNN-based fake news detectors typically consider the entities and relations that can represent a kind of structural pattern to construct a graph. As there are abundant structural patterns in the social context, the graph constructed by different detectors encompasses a variety of types, e.g., the user-news bipartite graph [Wang *et al.*, 2023a], the news engagement graph [Wu and Hooi, 2023], the news-user propagation tree [Dou *et al.*, 2021], and the news-post propagation and dispersion tree [Bian *et al.*, 2020]. For simplicity, we describe the graph construction process as:

$$G_s = r(G), \tag{1}$$

where $r(\cdot)$ is a function to generate a graph based on social context. After constructing the graph, detectors will utilize a pre-trained language model, e.g., GloVe [Pennington *et al.*, 2014] and BERT [Devlin *et al.*, 2019], to generate the feature matrix X and employ a GNN to learn news embedding, aiming to classify news.

Adversarial Attack on Specific Graph. Existing adversarial attack methods can easily perturb the graph by adding a small perturbation ξ to the graph G_s. Here, the attack process can be described as:

$$G_s' = G_s + \xi. \tag{2}$$

However, different detectors employ different functions to generate the graph, and thus attackers need to deal with a set of Graphs $\{G_s^1, \ldots, G_s^k\}$ where k is the number of the detectors. Due to structural differences, perturbation ξ is not general for the graphs in the set. In a black-box manner, the problem will become more challenging, thereby encouraging us to develop a general approach to attack the detectors.

Our Attack on Social Context. To execute a generalized attack against GNN-based fake news detectors, a natural way is to perturb the social context G instead of G_s. Therefore, a general attack process can be described as:

$$G' = G + \xi. \tag{3}$$

However, the small perturbation ξ may disrupt the semantic integrity, making attacks more easily detectable. Hence, we attempt to simulate social interaction to meet practical constraints. Specifically, we manipulate a set of controllable fraudsters U_c to add social interaction records through sharing. The perturbed social context can be represented as:

$$G' = (U, Q, P \cup P', E \cup E_u' \cup E_t'), \tag{4}$$

where P' includes new posts sent by fraudsters, E_u' and E_t' are new user-post and shared relations, respectively. The attack's objective for the k-th news q_k is formalized as:

$$\begin{aligned} &\max_{P', E'} \mathbb{I}(\arg\max_z f_{\theta^*}(r(G'), X)_{q_k} \neq y_{q_k}), \\ &\text{s.t. } \theta^* = \arg\min_\theta L(\theta; r(G), X), \\ &|P'| = |E_u'| = |E_t'| \leq \Delta, \end{aligned} \tag{5}$$

where f_θ is the targeted GNN-based fake new detector which has been trained on the clean data, $\mathbb{I}(\cdot)$ is an indicator function, X is the feature matrix, z is the predicted label of news q_k, and Δ is the given budget. The goal is to maximize the likelihood of q_k being misclassified within the constraints.

4 Method

In this section, we describe our general attack framework which consists of two major modules, i.e., the fraudster selection module and the post injection module, as shown in Figure 2. Given the social context of news dissemination G as input, our framework adds a set of social interaction records and eventually fools the GNN-based detectors.

4.1 Fraudster Selection Module

To select users to engage in the attack, following [Dou *et al.*, 2021; Wang *et al.*, 2023a], we first utilize a pre-trained language model [Pennington *et al.*, 2014] to extract user representation from the users' historical posts. However, the GNN-based detector's attention on the user representation is often different from the local structure and global structure. To comprehensively consider the impact of user representation on different structures, we utilize the propagation tree to estimate the local influence of the user representation and construct the user-news bipartite graph to estimate the global influence. Accordingly, we propose a user selection module that can effectively optimize the set of potential fraudsters.

Local Influence Estimation. To capture local influence, we first construct a propagation tree $G_t^k = (\mathcal{V}_t^k, E_t^k)$ for target news q_k, which consists of:

$$\begin{aligned} \mathcal{V}_t^k &= \{v | \forall v \in \text{traversal } (E_t, q_k)\}, \\ E_t^k &= \{(v_i, v_j) | \forall (v_i, v_j) \in \text{traversalpath } (E_t, q_k)\}, \end{aligned} \tag{6}$$

where traversal(E_t, q_k) indicates a function to traverse a tree with news q_k as the root according to E_t, and traversalpath(E_t, q_k) indicates a function to obtain the paths in above traverse process. Then, we let the representations of post nodes in G_t^k equal to the user representation of the post owner such that the GNN-based detector can learn user representation in a local view. Empirically, the gradient can characterize the model's attention on different inputs. Inspired by gradient-based attribution methods [Shrikumar *et al.*, 2017; Sanchez-Lengeling *et al.*, 2020], we leverage the model's attention on different feature elements to measure the local

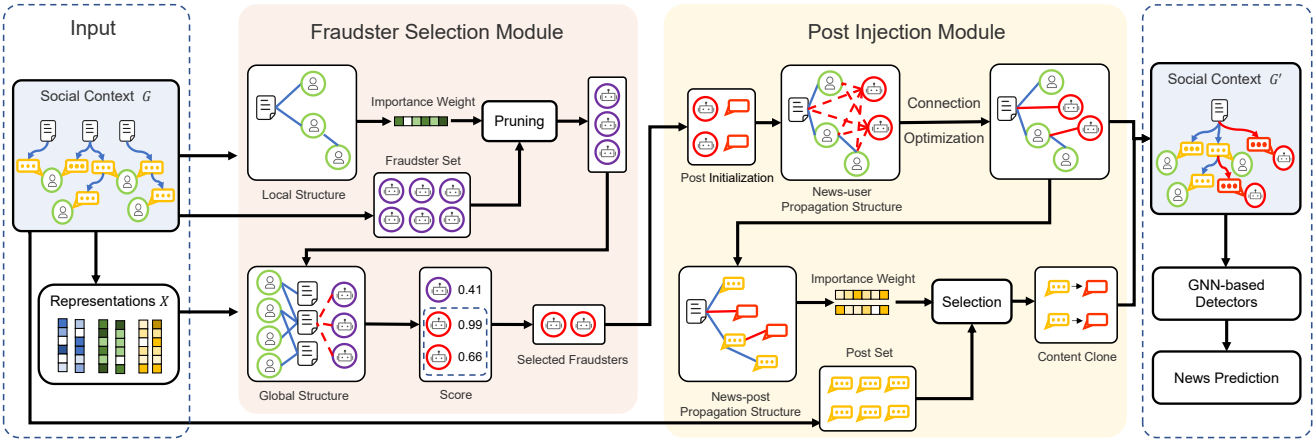


Figure 2: An illustration of GAFSI against GNN-based fake news detectors. The social context G and corresponding representation X serve as the input of GAFSI. The fraudster selection module selects influential fraudsters according to information from the local structure and global structure. Each selected fraudster will send a post and then the post injection module will optimize the connection and the content of the post. Finally, the records of sharing will be added into the social context and fool the GNN-based detector.

influence. Following the black-box manner, a GNN-based graph classifier f_{θ_t} is trained as a surrogate model to obtain attention information. Given the propagation tree G_t^k, the importance weight of the user representation is given by:

$$w_{G_t^k} = \frac{1}{N-1} \sum_{v_i \in \mathcal{V}_t^k \setminus \{q_k\}} \frac{\partial f_{\theta_t}(A_{G_t^k}, X_{G_t^k})_{1-y_{q_k}}}{\partial x_{v_i}}, \tag{7}$$

where $A_{G_t^k}$ is the adjacency matrix of G_t^k, $X_{G_t^k}$ is the user representation matrix of G_t^k, x_{v_i} is the user representation of the publisher of post v_i and N is the scale of $\mathcal{V}_t^k$. The j-th entry in $w_{G_t^k}$ indicates the relative importance for j-th feature element resulting from classifying q_k into the wrong class based on G_t^k. Intuitively, users with greater influence should have larger weights on the relatively important feature elements. According to the importance weight, the local influence of a fraudster $u_i \in U_c$ is measured as:

$$s_{u_i} = x_{u_i} \cdot w_{G_t^k}, \tag{8}$$

where x_{u_i} is the user representation of the fraudster u_i. To improve efficiency, we utilize local influence for pruning. Obviously, nodes with lower local influence are not ideal choices. Hence, $ceil(\alpha\Delta)$ fraudsters will be remained, where their local influence are $ceil(\alpha\Delta)$-largest.

Global Influence Estimation. To estimate the global influence of the user representation, we construct a user-news bipartite graph $G_b = (\mathcal{V}_b, E_b)$:

$$\begin{aligned} \mathcal{V}_b =& U \cup Q, \\ E_b =& \{(q_k, u_i) | \exists (u_i, p_j) \in E_u \wedge p_j \in \mathcal{V}_t^k\}. \end{aligned} \tag{9}$$

Due to the direct connections between users and news, we can utilize the gradient of the edge between them to estimate the global influence. Similarly, we train a GNN-based node classifier f_{θ_b} as the surrogate model based on G_b. Subsequently, the gradient of the potential edge between a fraudster u_i and the target news q_k is calculated as:

$$\nabla_{a_{q_k,u_i}} = \frac{\partial L(f_{\theta_b}(A_{G_b}, X_{G_b}))_{q_k}}{\partial a_{q_k,u_i}}, \tag{10}$$

where A_{G_b} is the adjacency matrix of G_b, X_{G_b} is the feature matrix of G_b and L is the cross-entropy loss function. In general, a larger gradient of the potential edge indicates that the corresponding user's participation in the attack will cause a greater impact on the target news. Hence, we select the fraudster u_i with the largest gradient in those remaining fraudsters and update the selected fraudster set $U_a = U_a \cup \{u_i\}$. The related edge is added to E_b and the selection process will be repeated until Δ fraudsters have been chosen.

4.2 Post Injection Module

Once a set of users U_a engaging in the attack is selected, we attempt to determine the targets of sharing behavior for these users such that the shared relations can be established.

Post Initialization. Since the budget for new interactions and the number of selected users are both Δ, each user $u_i \in U_a$ can send a new post to maximize user participation within the attack budget. Hence, we can obtain a set of new post P' and a set of user-post relation E'_u. Note that the content of the post is still hard to generate and we temporarily set it to empty. Intuitively, directly sharing the target news q_k is not always the best choice. To measure the impact of selecting different shared targets, we employ the previous propagation tree G_t^k to model the existing propagation structure of the news q_k. Then, the sharing behaviors can be treated as injecting posts into G_t^k such that we obtain a new propagation tree $G_o = (\mathcal{V}_o, E_o)$, where $\mathcal{V}_o = \mathcal{V}_t^k \cup P'$ and $E_o = E_t^k$. The new adjacency matrix of A_{G_o} can be formulated as:

$$A_{G_o} = \begin{bmatrix} A_{G_t^k} & B \\ B^T & O \end{bmatrix}, \tag{11}$$

where $A_{G_t^k}$ is the original adjacency matrix of G_t^k, B is the potential edges between injected posts and the posts in the

original tree and O is the potential edges between injected posts. At the beginning, B and O are both zero matrices.

Connection Optimization. Subsequently, we leverage the gradient of the edges in B as the proxy of attention to find the optimal connection. Since the content of the post has not been optimized, it is infeasible to optimize the connection based on the text representation. Instead, we utilize the user representation to select the shared target. According to E_u and E'_u, we let the representations of post nodes in G_o equal to the user representation of the post publisher. Then, a GNN-based graph classifier f_{θ_t} is utilized as a surrogate model to obtain gradient information. Different from the previous process, we only consider Δ injected posts. Hence, the simplified optimization process is described as:

$$\arg\max_{a_{ij}} \nabla_B \odot C = \arg\max_{a_{ij}} \frac{\partial L(f_{\theta_t}(A_{G_o}, X_{G_o}))}{\partial B} \odot C, \tag{12}$$

where $\odot$ represents the element-wise product, A_{G_o} is the adjacency matrix of G_o, X_{G_o} is the user representation matrix of G_o, C is a mask matrix and a_{ij} is the edge with the largest gradient under constraints in B. At the beginning, C is a matrix of ones. We treat a_{ij} represents an optimal shared relation and expand $E_o = E_o \cup \{(v_i, v_j)\}$. Each post has a unique source in the tree. To ensure each new post can only be selected once, we let $C_{kj} = 0, \forall k$. Iteratively, we expand E_o until each injected post has one connection in E_o. Hence, we obtain the set of new shared relation $E'_t = E_o \setminus E^k_t$.

Content Clone. According to the new propagation tree G_o, we optimize the text representation for injected posts which has been set as empty before. Recent study [Wang *et al.*, 2023b] directly set the feature as zero, indicating an empty message, leading to little impact on detectors. To address this problem, we attempt to clone existing content into our new posts. Similarly, we leverage the model's attention on the text representations to estimate the influence of the existing content. Specifically, we extract text representations from the content of existing posts and retrain a GNN-based graph classifier f'_{θ_t} according to the text representations. Then, for a new post $p_i \in P'$, we let the feature vector of it equal to the zero feature vector due to its empty content. Given the modified propagation tree G_o, the attention weight of the text representation for the new post p_i can be calculated as:

$$w_{p_i} = \frac{\partial f'_{\theta_t}(A_{G_o}, X'_{G_o})_{1-y_{q_k}}}{\partial x_{p_i}}, \tag{13}$$

where X'_{G_o} represents the text representation matrix of G_o, x_{p_i} is the feature vector for post p_i. Then, we randomly sample a set of potential post $P_r \subset P$. For each $p_j \in P_r$, we calculate the score as:

$$s_{p_{i,j}} = x_{p_j} \cdot w_{p_i}, \tag{14}$$

where x_{p_j} is the text representation of the post p_j. After computing the score for all posts in P_r, we select the potential post that has the largest score and clone the post content to the new post p_i. The content of all injected posts will be updated and then the posts can be injected into the social context.

Based on the two modules, we mix the generated records of sharing behavior into the raw data and then fool the GNN-based fake news detectors.

5 Experiment

5.1 Experimental Settings

Datasets. We adopt two real-world datasets [Shu *et al.*, 2017; Fey and Lenssen, 2019], i.e., Politifact and Gossipcop, from the PyTorch-Geometric library. To train detectors and our surrogate model, we split the data into 20% for the training, 10% for the validation, and 70% for the testing. The testing set is also treated as the set of the news to be attacked.

GNN-based Fake News Detectors. We investigate the robustness of different GNN-based fake news detectors. For social-context-based detectors, following [Wang *et al.*, 2023a], we construct a user-news bipartite graph. Leveraging Glove [Pennington *et al.*, 2014], we extract the features of user nodes from their historical posts and the features of news nodes from the news contents. Subsequently, we attack three variants of GNN, i.e., GCN [Kipf and Welling, 2017], GraphSAGE [Hamilton *et al.*, 2017], and GAT [Veličković *et al.*, 2018]. Besides, following [Wu and Hooi, 2023], we construct a news engagement graph and attack the related proposed model. For propagation-based detectors, following [Dou *et al.*, 2021], we construct a tree-structured propagation graph. The features of post nodes are derived from the historical posts of their owner. Then, we investigate three variants of UPFD [Dou *et al.*, 2021]. We also investigate BiGCN [Bian *et al.*, 2020] on a propagation and dispersion bi-directional tree. Different from the previous tree, the features of post nodes are derived from the post content.

For clarity, we use G1, G2, G3, and G4 to represent a bipartite graph, a news engagement graph, a propagation tree, and a bi-directional tree mentioned above, respectively.

Baselines. Since there is no attempt to conduct a general attack on detectors based on different graph structures, we need to make some necessary assumptions to extend the existing methods. The baseline method will conduct the attack on the user-news bipartite graph and connect users with the target news. To obtain a complete record of social interaction, we assume the users selected by the baselines randomly share the related post of target news and the new posts have empty content. The detailed baselines are described as follows:

Random, which randomly establishes a connection between a user and the target news.

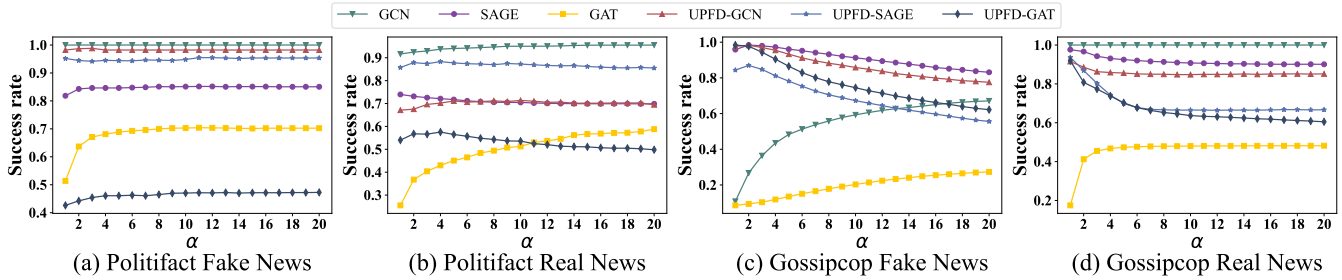
DICE [Waniek *et al.*, 2018], which randomly connects the target news to several users with a different label. The pseudo-labels of users are obtained from the surrogate model.

SGA [Li *et al.*, 2021], one of the most effective targeted attacks on the node classification task. SGA constructs a subgraph for the target node and computes gradients of edges in the subgraph to generate edge perturbation.

MARL [Wang *et al.*, 2023a], which is the first work to attack GNN-based fake news detectors utilizing a multiagent reinforcement learning framework.

Dataset	Graph	Model	Fake News						Real News					
			-	Random	DICE	MARL	SGA	GAFSI	-	Random	DICE	MARL	SGA	GAFSI
Politifact	G1	GCN	0.20	0.16	0.29	0.97	**1.00**	**1.00**	0.04	0.18	0.30	0.81	**0.95**	**0.95**
		SAGE	0.27	0.30	0.37	0.56	0.84	**0.85**	0.12	0.12	0.15	0.32	0.65	**0.70**
		GAT	0.21	0.24	0.34	0.43	0.68	**0.70**	0.10	0.12	0.18	0.39	**0.67**	0.55
	G2	DECOR-GCN	0.18	0.31	0.45	0.57	0.86	**0.92**	0.10	0.11	0.21	0.36	0.32	**0.37**
	G3	UPFD-GCN	0.26	0.29	0.31	0.39	0.55	**0.98**	0.10	0.11	0.11	0.13	0.15	**0.72**
		UPFD-SAGE	0.26	0.35	0.52	0.56	0.89	**0.96**	0.13	0.16	0.26	0.32	0.73	**0.86**
		UPFD-GAT	0.27	0.26	0.28	0.31	0.45	**0.48**	0.14	0.16	0.17	0.25	0.41	**0.52**
	G4	BiGCN	0.21	0.22	0.23	0.29	0.31	**0.91**	0.10	0.12	0.15	0.13	0.18	**0.92**
Gossipcop	G1	GCN	0.02	0.01	0.04	0.63	**0.77**	0.57	0.08	0.15	0.18	0.75	**1.00**	**1.00**
		SAGE	0.10	0.09	0.10	0.31	0.37	**0.92**	0.02	0.12	0.13	0.35	0.85	**0.91**
		GAT	0.04	0.05	0.08	0.21	**0.40**	0.23	0.03	0.08	0.11	0.22	**0.48**	**0.48**
	G2	DECOR-GCN	0.07	0.06	0.09	0.23	0.64	**0.78**	0.04	0.13	0.15	0.08	0.20	**0.24**
	G3	UPFD-GCN	0.02	0.02	0.04	0.05	0.06	**0.85**	0.05	0.06	0.08	0.08	0.12	**0.84**
		UPFD-SAGE	0.02	0.03	0.06	0.04	0.21	**0.68**	0.05	0.05	0.07	0.18	0.52	**0.68**
		UPFD-GAT	0.02	0.03	0.06	0.04	0.20	**0.75**	0.06	0.06	0.08	0.16	0.44	**0.64**
	G4	BiGCN	0.04	0.05	0.07	0.05	0.08	**0.63**	0.05	0.07	0.08	0.09	0.10	**0.78**

Table 1: The success rate of GAFSI compared to other baselines. Fake News and Real News represent the original label of the target news. "-" denotes the misclassification rate for the detectors before the attack. The best results are highlighted in bold.



(a) Politifact Fake News (b) Politifact Real News (c) Gossipcop Fake News (d) Gossipcop Real News

Figure 3: The success rate of GAFSI when adopting different trade-off parameters α.

5.2 Performance Analysis

Comparison with State-of-the-art Methods. Our goal is to change the prediction of the target news. To this end, we evaluate the success rate of the attack which is defined as the number of misclassified news after the attack divided by the total number of target news. A larger success rate means better attack performance. Following the existing targeted attacks [Zügner *et al.*, 2018; Li *et al.*, 2021], the budget of new records Δ equals to the degree of the target node in the user-news bipartite graph which represents the original number of people engaged in sharing the news. For each target news, we repeat the attack 10 times and report their average results.

Table 1 shows the attack performance when changing the prediction of fake news and real news respectively. Our method outperforms all baselines in most cases, particularly against propagation-based detectors on G3 and G4. Besides, GAFSI achieves a greater improved amount of success rate in the larger dataset, i.e., Gossipcop, demonstrating the ability to deal with large-scale graphs. MARL fails to reduce the search space, achieving poor attack performance. Compared to SGA, GAFSI has less influence on target news in three cases. This is because SGA solely focuses on modifying the user-news bipartite graph, while GAFSI makes a trade-off between different graph structures, improving the transferability with relatively minor costs. Furthermore, the attack performance between different variants also improves in some cases, especially against GraphSAGE.

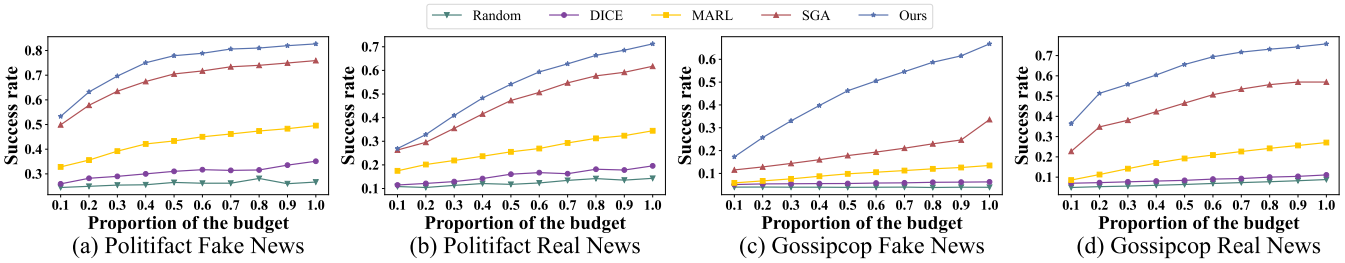
Dataset	MARL	SGA	GAFSI
Politifact	51.77	0.78	**0.50**
Gossipcop	54.24	2.20	**1.69**

Table 2: The average running time (s) of attacking a piece of news.

Computational Efficiency. We demonstrate the average running time for attacking single news in Table 2. Compared to SGA, we generate more complex and complete social interaction in a similar running time. MARL suffers from low time efficiency due to the vast search space and multiple samplings. In contrast to MARL, our approach shows a noticeable advantage in computational efficiency through utilizing a hierarchical strategy in the fraudster selection module.

5.3 Ablation Studies and Other Analysis

Effects of Varying Hyper-parameters. We examine the performance of GAFSI with different parameters α to make a trade-off between local influence and global influence in the fraudster selection module. The results against detectors on G1 and G3 are shown in Figure 3, GAFSI achieves a poor performance against detectors on G1 when only considering local influence, i.e., $\alpha = 1$. As α increases, we observe that enhancing the attack performance against detectors on



(a) Politifact Fake News (b) Politifact Real News (c) Gossipcop Fake News (d) Gossipcop Real News

Figure 4: The average success rate of attacks with different budgets.

Dataset	Model	-	Root	Random	Ours
Politifact	UPFD-GCN	0.17	0.79	0.37	**0.84**
	UPFD-SAGE	0.19	0.81	0.86	**0.90**
	UPFD-GAT	0.20	0.48	0.48	**0.50**
Gossipcop	UPFD-GCN	0.04	0.82	0.39	**0.84**
	UPFD-SAGE	0.03	0.63	0.61	**0.68**
	UPFD-GAT	0.04	0.56	0.62	**0.70**

Table 3: The success rate for different post sources.

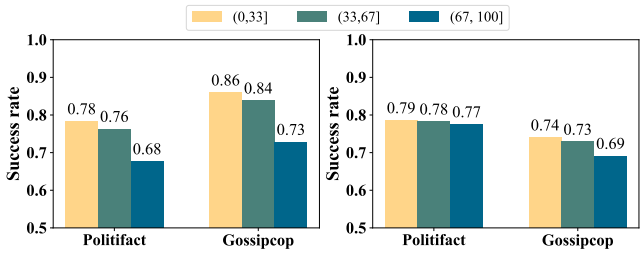
Dataset	Model	-	Empty	Random	Ours
Politifact	UPFD-GCN	0.10	0.13	0.19	**0.95**
	UPFD-SAGE	0.11	0.13	0.14	**0.68**
	UPFD-GAT	0.11	0.12	0.12	**0.48**
	BiGCN	0.12	0.14	0.20	**0.91**
Gossipcop	UPFD-GCN	0.03	0.04	0.17	**0.70**
	UPFD-SAGE	0.02	0.02	0.20	**0.27**
	UPFD-GAT	0.02	0.02	0.25	**0.39**
	BiGCN	0.04	0.12	0.12	**0.71**

Table 4: The success rate for different post content.

G1 sacrifices the attack performance against detectors on G3. When α is smaller, this trade-off is worthwhile. Ultimately, the method's attack performance achieves a balance, indicating that subsequent nodes have a low impact on any models.

Effects of Different Post Sources. To demonstrate the validity of the post injection module in GAFSI, we conduct ablation experiments on the connection strategy. A source of a post could be a news or a related post. To conduct the ablation study, we provide two strategies for comparison. Fraudsters can choose to directly share the target news, which is regarded as a Root strategy. Another choice for fraudsters is to randomly share a related post, which is regarded as a Random strategy. We conduct experiments against detectors on G3. As shown in Table 3, we compare these two strategies to ours, indicating that our approach can cause a greater impact on the propagation tree of the target news.

Effects of Different Post Content. For the ablation study of post content generation, we directly set the new post as an empty text or randomly assign the existing content to it. In contrast to previous experiments, the detectors to be attacked classify news based on the content of posts. Therefore, we leverage text representation to train the detector on G3 and G4 before the attack. Table 4 reports the results of different strategies against detectors that leverage text representation as post nodes' features. It is obvious that our strategy contributes to enhancing the attack performance.



(a) Detectors on G1 (b) Detectors on G3

Figure 5: The success rate of GAFSI on target news within different degree ranges.

Effects of Different Budgets. To better evaluate the effectiveness of GAFSI with different budgets, as shown in Figure 4, we gradually reduce the budget to observe changes in attack performance on the detectors based on G1 and G3, respectively. As the budget increases, our method widens the gap in attack performance compared to other methods, especially for the fake news in Gossipcop. We also observe a slowing growth rate, which could be attributed to two factors. Firstly, later-selected users often have lower influences than earlier-selected ones, Secondly, the increase in the number of users might dilute the impact of single users.

Effects of Different News Degrees. Intuitively, the degree of node has a large impact on attack performance. Since relatively low degrees constitute a significant portion of real-world datasets, nodes that have a degree less than 100 are selected for further investigation. We group the target news based on their node degrees on the bipartite graph. Then, we summarize the attack results for the grouped nodes. As shown in Figure 5, the success rate of our method decreases when the degree of target news increases. This implies that the detectors are more robust when predicting news with more participants involved. Besides, compared to detectors working on G1, those working on G3 seem less sensitive facing varying degrees, which may be attributed to their tree structures.

6 Conclusion

In this paper, we propose a general black-box adversarial attack framework against GNN-based fake news detectors with different graph structures. The key idea is to add fake social interaction to the social context via fraudster's sharing behaviors. Extensive experimental results demonstrate the effectiveness and superiority of our method in attacking different fake news detectors. In the future, we plan to investigate the engagement pattern and realize a more imperceptible attack.

Acknowledgements

This work was supported in part by the National Science Fund for Distinguished Young Scholarship of China (Grant no. 62025602), the National Natural Science Foundation of China (Grant nos. U22B2036, 62073263, 62203363, 62102105), the Fundamental Research Funds for the Central Universities (Grant no. D5000230112), the Tencent Foundation and XPLORER PRIZE, the Young Talent Fund of Association for Science and Technology in Shaanxi (Grant no. 20240105), and the Shaanxi Provincial Natural Science Foundation (Grant no. 2024JC-YBQN-0620).

References

[Aïmeur *et al.*, 2023] Esma Aïmeur, Sabrine Amri, and Gilles Brassard. Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining*, 13:30, 2023.

[Bian *et al.*, 2020] Tian Bian, Xi Xiao, Tingyang Xu, Peilin Zhao, Wenbing Huang, Yu Rong, and Junzhou Huang. Rumor detection on social media with bi-directional graph convolutional networks. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, pages 549–556, 2020.

[Cheng *et al.*, 2024] Le Cheng, Peican Zhu, Keke Tang, Chao Gao, and Zhen Wang. Gin-sd: Source detection in graphs with incomplete nodes via positional encoding and attentive fusion. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, pages 55–63, 2024.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, 2019.

[Dong *et al.*, 2023] Yiqi Dong, Dongxiao He, Xiaobao Wang, Yawen Li, Xiaowen Su, and Di Jin. A generalized deep markov random fields framework for fake news detection. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence*, pages 4758–4765, 2023.

[Dong *et al.*, 2024] Yiqi Dong, Dongxiao He, Xiaobao Wang, Yawen Li, Xiaowen Su, and Di Jin. Unveiling implicit deceptive patterns in multi-modal fake news via neuro-symbolic reasoning. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, pages 8354–8362, 2024.

[Dou *et al.*, 2021] Yingtong Dou, Kai Shu, Congying Xia, Philip S. Yu, and Lichao Sun. User preference-aware fake news detection. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2051–2055, 2021.

[Fey and Lenssen, 2019] Matthias Fey and Jan Eric Lenssen. Fast graph representation learning with pytorch geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

[Guo *et al.*, 2024] Sensen Guo, Xiaoyu Li, Peican Zhu, Baocang Wang, Zhiying Mu, and Jinxiong Zhao. Mixcam-attack: Boosting the transferability of adversarial examples with targeted data augmentation. *Information Sciences*, 657:119918, 2024.

[Hamilton *et al.*, 2017] William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 1025–1035, 2017.

[Han *et al.*, 2021] Yi Han, Shanika Karunasekera, and Christopher Leckie. Continual learning for fake news detection from social media. In *International Conference on Artificial Neural Networks 2021*, pages 372–384, 2021.

[Hua *et al.*, 2023] Jiaheng Hua, Xiaodong Cui, Xianghua Li, Keke Tang, and Peican Zhu. Multimodal fake news detection through data augmentation-based contrastive learning. *Applied Soft Computing*, 136:110125, 2023.

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

[Li *et al.*, 2021] Jintang Li, Tao Xie, Liang Chen, Fenfang Xie, Xiangnan He, and Zibin Zheng. Adversarial attack on large scale graph. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):82–95, 2021.

[Liu *et al.*, 2022] Zihan Liu, Yun Luo, Lirong Wu, Zicheng Liu, and Stan Z. Li. Towards reasonable budget allocation in untargeted graph structure attacks via gradient debias. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pages 27966–27977, 2022.

[Nguyen *et al.*, 2020] Van-Hoang Nguyen, Kazunari Sugiyama, Preslav Nakov, and Min-Yen Kan. Fang: Leveraging social context for fake news detection using graph representation. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 1165–1174, 2020.

[Pennington *et al.*, 2014] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543, 2014.

[Phan *et al.*, 2023] Huyen Trang Phan, Ngoc Thanh Nguyen, and Dosam Hwang. Fake news detection: A survey of graph neural network methods. *Applied Soft Computing*, 139:110235, 2023.

[Sanchez-Lengeling *et al.*, 2020] Benjamin Sanchez-Lengeling, Jennifer Wei, Brian Lee, Emily Reif, Peter Y. Wang, Wesley Wei Qian, Kevin McCloskey, Lucy Colwell, and Alexander Wiltschko. Evaluating attribution for graph neural networks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 5898–5910, 2020.

[Shang *et al.*, 2023] Yu Shang, Yudong Zhang, Jiansheng Chen, Depeng Jin, and Yong Li. Transferable structure-based adversarial attack of heterogeneous graph neural network. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 2188–2197, 2023.

[Shrikumar *et al.*, 2017] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3145–3153, 2017.

[Shu *et al.*, 2017] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36, 2017.

[Silva *et al.*, 2021] Amila Silva, Yi Han, Ling Luo, Shanika Karunasekera, and Christopher Leckie. Propagation2vec: Embedding partial propagation networks for explainable fake news early detection. *Information Processing & Management*, 58(5):102618, 2021.

[Su *et al.*, 2023] Xing Su, Jian Yang, Jia Wu, and Yuchen Zhang. Mining user-aware multi-relations for fake news detection in large scale online social networks. In *Proceedings of the 16th ACM International Conference on Web Search and Data Mining*, pages 51–59, 2023.

[Tang *et al.*, 2023] Keke Tang, Yawen Shi, Tianrui Lou, Weilong Peng, Xu He, Peican Zhu, Zhaoquan Gu, and Zhihong Tian. Rethinking perturbation directions for imperceptible adversarial attacks on point clouds. *IEEE Internet of Things Journal*, 10(6):5158–5169, 2023.

[Tao *et al.*, 2021] Shuchang Tao, Qi Cao, Huawei Shen, Junjie Huang, Yunfan Wu, and Xueqi Cheng. Single node injection attack against graph neural networks. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 1794–1803, 2021.

[Veličković *et al.*, 2018] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *Proceedings of the 6th International Conference on Learning Representations*, 2018.

[Wang *et al.*, 2022] Zhengyi Wang, Zhongkai Hao, Ziqiao Wang, Hang Su, and Jun Zhu. Cluster attack: Query-based adversarial attacks on graph with graph-dependent priors. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence*, pages 768–775, 2022.

[Wang *et al.*, 2023a] Haoran Wang, Yingtong Dou, Canyu Chen, Lichao Sun, Philip S. Yu, and Kai Shu. Attacking fake news detectors via manipulating news social engagement. In *Proceedings of the ACM Web Conference 2023*, pages 3978–3986, 2023.

[Wang *et al.*, 2023b] Lanjun Wang, Xinran Qiao, Yanwei Xie, Weizhi Nie, Yongdong Zhang, and Anan Liu. My brother helps me: Node injection based adversarial attack on social bot detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6705–6714, 2023.

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

[Wu and Hooi, 2023] Jiaying Wu and Bryan Hooi. Decor: Degree-corrected social graph refinement for fake news detection. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2582–2593, 2023.

[Yin *et al.*, 2024] Shu Yin, Peican Zhu, Lianwei Wu, Chao Gao, and Zhen Wang. Gamc: An unsupervised method for fake news detection using graph autoencoder with masking. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, pages 347–355, 2024.

[Zhang *et al.*, 2023] Mengmei Zhang, Xiao Wang, Chuan Shi, Lingjuan Lyu, Tianchi Yang, and Junping Du. Minimum topology attacks for graph neural networks. In *Proceedings of the ACM Web Conference 2023*, pages 630–640, 2023.

[Zhu *et al.*, 2023] Peican Zhu, Jinbang Hong, Xingyu Li, Keke Tang, and Zhen Wang. Sgma: a novel adversarial attack approach with improved transferability. *Complex & Intelligent Systems*, 9:6051–6063, 2023.

[Zhu *et al.*, 2024] Peican Zhu, Zepeng Fan, Sensen Guo, Keke Tang, and Xingyu Li. Improving adversarial transferability through hybrid augmentation. *Computers & Security*, 139:103674, 2024.

[Zügner *et al.*, 2018] Daniel Zügner, Amir Akbarnejad, and Stephan Günnemann. Adversarial attacks on neural networks for graph data. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2847–2856, 2018.