Poisoning $\times$ Evasion: Symbiotic Adversarial Robustness for Graph Neural Networks

Ege Erdogan, Simon Geisler, Stephan Günnemann
{ege.erdogan, s.geisler, s.guennemann}@tum.de
Department of Computer Science
Technical University of Munich

Abstract

It is well-known that deep learning models are vulnerable w.r.t. small input perturbations. Such perturbed instances are called adversarial examples. Adversarial examples are commonly crafted to fool a model either at training time (poisoning) or test time (evasion). In this work, we study the symbiosis of poisoning and evasion. We show that combining both threat models can substantially improve the devastating efficacy of adversarial attacks. Specifically, we study the robustness of Graph Neural Networks (GNNs) under structure perturbations and devise a memory-efficient adaptive end-to-end attack for the novel threat model using first-order optimization.

1 Introduction

Graph neural networks (GNNs) are increasingly used in different domains such as product recommendations (Hao et al., 2020) and drug discovery (Guo et al., 2021). Nevertheless, GNNs are vulnerable to adversarial attacks across many tasks such as node classification (Zügner et al., 2018; Dai et al., 2018; Geisler et al., 2021), graph classification (Dai et al., 2018; Wang et al., 2023), link prediction (Chen et al., 2020), and node embeddings (Bojchevski and Günnemann, 2019; Zhang et al., 2019). With attacks being able to scale to very large graphs (Geisler et al., 2021), studying the adversarial robustness of GNNs is of growing importance. GNNs can be attacked during test time (evasion) or train time (poisoning); yet, a threat model combining evasion and poisoning has not been considered in the literature. It is nevertheless a reasonable threat model considering e.g. publicly available graphs, or graphs extracted from sources such as social media sites.

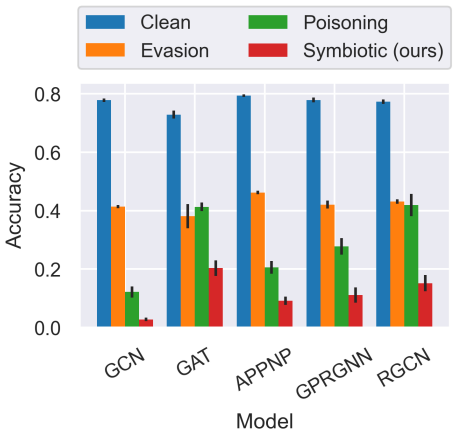

Figure 1: Perturbed accuracies (with standard errors) after evasion, poisoning, and symbiotic (ours) attacks on different models on PubMed.

The Problem. We consider node classification tasks, and an adversary able to change the structure of the graph (i.e. insert/remove edges), with both train and test time access to the graph. The adversary's end goal is to minimize the classification accuracy on the test set. It is constrained by a global budget and attacks entire graph at once rather than targeting specific nodes. We call such attacks combining evasion and poisoning in this threat model *symbiotic* attacks.

NeurIPS 2023 New Frontiers in Graph Learning Workshop (NeurIPS GLFrontiers 2023).

Contributions. We initiate the study of this threat model and compare it with plain poisoning and evasion adversaries by adapting the previous PR-BCD attack (Geisler et al., 2021) to the symbiotic threat model, resulting in memory-efficient attacks that can scale to large graphs. Our main findings are:

- The symbiotic attacks are consistently stronger than a poisoning attack, indicating that it is beneficial to allocate part of the resources of the poisoning attack for the evasion objective.
- Evasion attacks are constrained by the number of test nodes, with larger test sets making an evasion attack harder. The symbiotic attacks are affected less significantly by the size of the test set since they can also manipulate the graph during training, leading to almost-zero accuracy in cases such as when the share of labeled train is low and test nodes high.

Overall, the significance of the potential improvement the symbiotic threat model provides indicates that it requires further study, and we outline paths of future work in our conclusion.

2 Preliminaries

Notation. Throughout we denote a graph by $\mathcal{G}$ with n nodes, adjacency matrix $\mathbf{A} \in \{0,1\}^{n \times n}$ and feature matrix $\mathbf{X} \in \mathrm{R}^{n \times d}$, and a GNN applied to the graph as $f_\theta(\mathcal{G})$ with parameters θ. $\Phi(\mathcal{G})$ is the set of admissible adversarial graphs resulting from $\mathcal{G}$, and $\mathcal{L}_{\text{atk}}$ and $\mathcal{L}_{\text{train}}$ denote the adversarial and training objectives.

2.1 Adversarial Robustness of GNNs

An adversarial attack on a GNN can change the structure of the graph by inserting/removing edges and nodes, or can modify the node features. We focus on node classification and edge-level structural perturbations.

Attacks can be divided into two categories: *evasion* and *poisoning*. An evasion attack targets a fixed GNN with θ obtained on a clean graph, and thus tries to solve the optimization problem

$$\max_{\hat{\mathcal{G}} \in \Phi(\mathcal{G})} \mathcal{L}_{\text{atk}}(f_\theta(\hat{\mathcal{G}})), \tag{1}$$

while a poisoning attack is performed before training but aims to degrade performance after training:

$$\max_{\hat{\mathcal{G}} \in \Phi(\mathcal{G})} \mathcal{L}_{\text{atk}}(f_{\theta^*}(\hat{\mathcal{G}})) \quad \text{where} \quad \theta^* = \operatorname*{argmin}_\theta \mathcal{L}_{\text{train}}(f_\theta(\hat{\mathcal{G}})). \tag{2}$$

A poisoning attack is admittedly more challenging. Previous work has tried using evasion perturbations as poisoning perturbations (Zügner et al., 2018), or unrolling the training procedure as part of the optimization to compute meta-gradients (gradients w.r.t. hyperparameters) of $\mathcal{L}_{\text{atk}}$ w.r.t. $\mathbf{A}$ (Zügner et al., 2020).

Finally, since we only consider changes to the binary adjacency matrix, we define $\Phi(\mathcal{G})$ to include graphs reachable from $\mathcal{G}$ after at most Δ edge perturbations (e.g. for undirected $\mathcal{G}$, $\Phi(\mathcal{G}) = \{\tilde{\mathcal{G}} \mid \|\tilde{\mathbf{A}} - \mathbf{A}\|_0 \leq 2\Delta \wedge \tilde{\mathbf{A}}^\top = \tilde{\mathbf{A}}\}$), though it would also be straightforward to use metrics such as the graph's degree distribution or diameter.

2.1.1 PR-BCD

Our work builds on the *Projected Randomized Block Coordinate Descent* (PR-BCD) attack proposed in Geisler et al. (2021). Similar to the Projected Gradient Descent (PGD) attack (Xu et al., 2019), the adjacency matrix is relaxed to $\mathbf{P} \in [0,1]^{n \times n}$ to enable continuous gradient updates, and each entry denotes the probability of flipping that edge with the final perturbations sampled from Bernoulli($\mathbf{P}$). However, since the adjacency matrix grows quadratically with the number of nodes, scalability of plain PGD becomes a challenge on larger graphs.

PR-BCD employs Randomized Block Coordinate Descent (R-BCD) (Nesterov, 2012) and updates a block of size b of $\mathbf{P}$ in each iteration. The projection step ensures that the budget is enforced in expectation; i.e. $\mathrm{E}[\text{Bernoulli}(\mathbf{P})] = \sum \mathbf{P} \leq \Delta$ and that $\mathbf{P} \in [0,1]^{n \times n}$. After each iteration, rather than sampling the entire block again, the promising entries of the block are kept and only the rest is

resampled. In other words, $\mathbf{P}$ is kept sparse in a survival-of-the-fittest manner. Multiple samples are drawn at the end and the best-performing one is returned as the final perturbed graph.

PGD can also be applied for a poisoning attack as the Meta-PGD attack in Mujkanovic et al. (2022). In our attacks, we employ the same principle with PR-BCD to scale better to larger graphs.

While we only consider a single global budget Δ, it is straightforward to include more sophisticated constraints when beneficial for the application at hand (Gosch et al., 2023b,a).

3 Symbiotic Attacks

The Symbiotic Objective. The problem of a symbiotic attack has a similar form to the bi-level optimization of Equation 2, but the main objective is conditioned on the evasion graph $\mathcal{G}^*$ in addition to the parameters θ^*:

$$\begin{aligned} \max_{\hat{\mathcal{G}} \in \Phi(\mathcal{G})} \mathcal{L}_{\text{pois}}(f_{\theta^*}(\mathcal{G}^*)) \quad \text{where} \quad \theta^* = \operatorname*{argmin}_{\theta} \mathcal{L}_{\text{train}}(f_\theta(\hat{\mathcal{G}})), \\ \text{and} \quad \mathcal{G}^* = \operatorname*{argmax}_{\tilde{\mathcal{G}} \in \Phi(\hat{\mathcal{G}})} \mathcal{L}_{\text{ev}}(f_{\theta^*}(\tilde{\mathcal{G}})) \end{aligned} \tag{3}$$

where for clarity we separate the poisoning and evasion objectives $\mathcal{L}_{\text{pois}}$ and $\mathcal{L}_{\text{ev}}$ although they might be the same function.

Threat Model. We model an attacker with the goal of degrading a model's performance on node classification tasks. Combining poisoning and evasion adversaries, our attacker has full access to the graph both at train and test times, knows the model architecture being used so that it can create surrogate models, but can access the trained model only as a black-box. Finally, our attacker is limited by a global budget of edge insertions/removals.

With the symbiotic objective and the threat model, we propose two attacks as approximations to the optimal solution.

The Sequential Attack. A simple way of launching a symbiotic attack is to split the budget into two, and then to launch an evasion attack with the second part after poisoning the model using the first part. In this case, the poisoning attack has no knowledge of a future evasion, but it can can still help the evasion attack by causing some nodes to be misclassified and reducing the classification margin of some nodes, making them easier to be misclassified during an evasion attack.

The Joint Attack. The poisoning attack can also be designed to "fit" the future evasion graph by including the evasion attack within the poisoning loss. In this *joint* attack, an evasion attack is launched within each poisoning attack iteration, and the poisoning loss is computed with the poisoned model over the evasion graph, so that the poisoning attack not only reduces the model's accuracy but also makes it more vulnerable to evasion.

Although the joint and sequential attacks can be instantiated using different evasion/poisoning attacks, we choose to build on PR-BCD since it scales well to larger graphs. This is a significant consideration since especially for the joint attack the inner evasion attack has to performed many times. Also note that the sequential attack is in fact a special case of the joint attack with zero iterations per inner evasion attack.

4 Evaluation

4.1 Setup

In this section, we compare the symbiotic threat model with evasion and poisoning attacks, instantiated through the PR-BCD evasion (Geisler et al., 2021) and poisoning attacks (Mujkanovic et al., 2022) on Cora, CiteSeer (McCallum et al., 2000), and PubMed (Sen et al., 2008). We study the robustness of GCN (Kipf and Welling, 2016), GAT (Veličković et al., 2018), APPNP (Gasteiger et al., 2018), and

Table 1: Numbers of nodes, edges, and classes in the datasets we include in our evaluations.

Dataset	Nodes	Edges	Classes
Cora	2,708	10,556	7
CiteSeer	3,327	9,104	6
PubMed	19,717	88,648	3

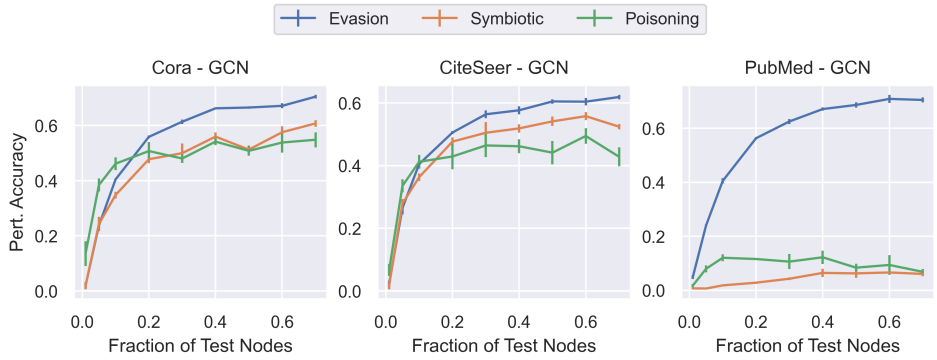

Figure 2: Perturbed accuracies and standard errors after the four attacks with different test set sizes and a fixed 5% global budget, using a GCN on the three benchmark datasets.

Figure 3: Classification accuracy after the four attacks on a GCN, with the varying **block sizes** for the PR-BCD optimization displayed along the x-axis.

GPRGNN (Chien et al., 2020). We also consider R-GCN (Zhu et al., 2019) and Jaccard purification (Wu et al., 2019) as potential defensive mechanisms. For each dataset, we allocate 20 nodes of each class for the labeled training set and allocate 10% of the nodes as the test set. For transductive tasks, the test nodes are also included as unlabeled train nodes during training, whereas for inductive tasks the test nodes are added to the graph after training. Table 1 displays the numbers of nodes, edges, and classes in each of our benchmark datasets.

4.2 Results

Table 2 displays the perturbed accuracy values on the test set (10% of nodes) for our benchmark datasets/models averaged over 10 runs, along with the standard error of the mean. We limit the attacker with a budget 5% the number of edges, with the budget split equally between poisoning and evasion for the symbiotic attacks. To better focus on the threat model, we report the better-performing of the two attacks for the symbiotic threat model and leave a comparison between the attacks to the appendix.

Symbiotic attacks are stronger than poisoning across all tasks, and stronger than plain evasion especially against GCN and GPRGNN models. The effect of the symbiotic threat model is most evident on the larger PubMed graph, with the accuracy dropping to almost zero e.g. against a GCN, especially as the evasion attack's performance drops against larger test sets as we discuss next.

Table 2: Average (± standard error) perturbed accuracies for the evasion, poisoning, and symbiotic attacks with a 5% budget. The -J suffix indicates the graph has been pre-processed with Jaccard purification (Wu et al., 2019) and *(ind.)* stands for inductive learning. The strongest (lowest accuracy) results for each setup are written in bold.

Model	Dataset	Clean	Evasion	Poisoning	Symbiotic
GCN	CiteSeer	0.68 ± 0.01	0.41 ± 0.01	0.4 ± 0.01	**0.38 ± 0.01**
	CiteSeer (ind.)	0.67 ± 0.01	0.41 ± 0.01	0.62 ± 0.01	**0.33 ± 0.01**
	CiteSeer-J	0.68 ± 0.01	0.41 ± 0.01	0.41 ± 0.02	**0.38 ± 0.01**
	Cora	0.78 ± 0.01	0.41 ± 0.01	0.46 ± 0.02	**0.35 ± 0.01**
	Cora (ind.)	0.75 ± 0.02	0.42 ± 0.01	0.68 ± 0.03	**0.3 ± 0.01**
	Cora-J	0.74 ± 0.01	0.39 ± 0.01	0.43 ± 0.02	**0.36 ± 0.01**
	PubMed	0.78 ± 0.01	0.41 ± 0.01	0.12 ± 0.02	**0.03 ± 0.01**
	PubMed-J	0.77 ± 0.01	0.41 ± 0.01	0.11 ± 0.01	**0.02 ± 0.0**
GAT	CiteSeer	0.62 ± 0.02	**0.27 ± 0.02**	0.41 ± 0.02	0.3 ± 0.03
	CiteSeer (ind.)	0.68 ± 0.01	**0.37 ± 0.01**	0.64 ± 0.02	0.56 ± 0.02
	CiteSeer-J	0.64 ± 0.01	0.32 ± 0.03	0.41 ± 0.03	**0.3 ± 0.03**
	Cora	0.69 ± 0.02	**0.22 ± 0.02**	0.48 ± 0.03	0.29 ± 0.02
	Cora (ind.)	0.77 ± 0.01	**0.21 ± 0.01**	0.61 ± 0.04	0.35 ± 0.03
	Cora-J	0.67 ± 0.01	**0.23 ± 0.02**	0.45 ± 0.02	0.28 ± 0.02
	PubMed	0.73 ± 0.01	0.38 ± 0.04	0.41 ± 0.01	**0.2 ± 0.03**
	PubMed-J	0.74 ± 0.01	0.34 ± 0.04	0.38 ± 0.04	**0.19 ± 0.02**
APPNP	CiteSeer	0.69 ± 0.01	**0.45 ± 0.01**	0.56 ± 0.01	0.47 ± 0.01
	CiteSeer (ind.)	0.71 ± 0.01	0.47 ± 0.01	0.66 ± 0.02	**0.4 ± 0.01**
	CiteSeer-J	0.68 ± 0.01	**0.43 ± 0.01**	0.52 ± 0.02	0.45 ± 0.02
	Cora	0.82 ± 0.02	**0.48 ± 0.03**	0.64 ± 0.02	0.51 ± 0.04
	Cora (ind.)	0.82 ± 0.02	0.53 ± 0.02	0.78 ± 0.01	**0.37 ± 0.01**
	Cora-J	0.82 ± 0.01	**0.5 ± 0.01**	0.67 ± 0.01	0.54 ± 0.01
	PubMed	0.79 ± 0.0	0.46 ± 0.01	0.21 ± 0.02	**0.09 ± 0.01**
	PubMed-J	0.77 ± 0.01	0.45 ± 0.01	0.19 ± 0.03	**0.1 ± 0.02**
GPRGNN	CiteSeer	0.66 ± 0.01	0.34 ± 0.01	0.44 ± 0.02	**0.33 ± 0.01**
	CiteSeer (ind.)	0.67 ± 0.01	0.37 ± 0.01	0.56 ± 0.01	**0.34 ± 0.01**
	CiteSeer-J	0.65 ± 0.01	**0.35 ± 0.01**	0.44 ± 0.01	**0.35 ± 0.01**
	Cora	0.82 ± 0.01	0.46 ± 0.01	0.53 ± 0.01	**0.4 ± 0.01**
	Cora (ind.)	0.8 ± 0.02	0.44 ± 0.01	0.74 ± 0.01	**0.35 ± 0.01**
	Cora-J	0.79 ± 0.01	0.44 ± 0.01	0.54 ± 0.01	**0.4 ± 0.01**
	PubMed	0.78 ± 0.01	0.42 ± 0.01	0.28 ± 0.03	**0.08 ± 0.02**
	PubMed-J	0.78 ± 0.01	0.42 ± 0.01	0.38 ± 0.04	**0.15 ± 0.04**
RGCN	CiteSeer	0.63 ± 0.01	**0.39 ± 0.01**	0.59 ± 0.02	0.47 ± 0.01
	Cora	0.74 ± 0.02	**0.44 ± 0.01**	0.74 ± 0.01	0.52 ± 0.02
	PubMed	0.77 ± 0.01	0.43 ± 0.01	0.42 ± 0.04	**0.15 ± 0.03**

4.3 Effect of the Number of Test Nodes

To illustrate a fundamental point of difference between poisoning and evasion objectives, Figure 2 displays the perturbed accuracies for evasion, poisoning, and symbiotic attacks over varying fractions of test nodes using a GCN with a 5% global budget. Attacking a very small number of nodes is easy as all attacks can obtain zero accuracy since a few edge perturbations should be enough to manipulate a small number of nodes.

As the number of test nodes grows, evasion becomes considerably more difficult across all datasets. Although poisoning and symbiotic attacks become more difficult as well with more test nodes, especially on PubMed they are much more robust than the evasion attack. Thus the degrading performance cannot be explained only by the attacks having to use the same budget to target a larger number of nodes. The poisoning attack is more lightly affected as it can also manipulate the flow of

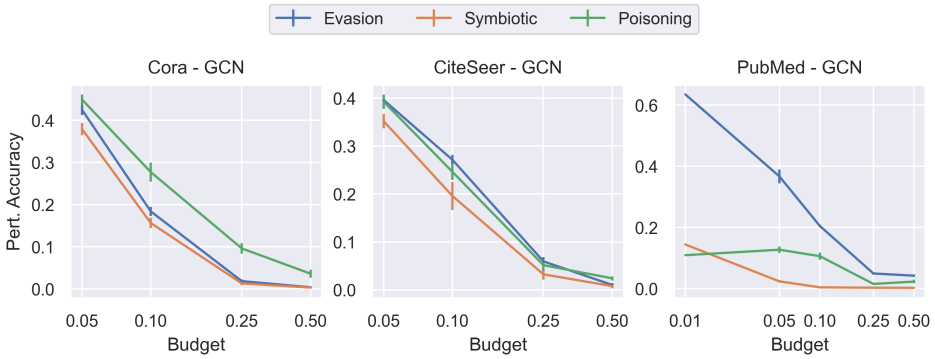

Figure 4: Perturbed accuracy of a GCN after the four attacks. The x-axis shows the global **budget** as a fraction of the number of edges.

information across the graph during training; e.g. the PubMed in our setup has the lowest share of labeled train nodes (20 nodes per class), which leads to poisoning attacks having devastating impact.

Similarly for the symbiotic attacks, poisoning helps in two ways: by reducing the base accuracy evasion starts with, and possibly changing the structure of the graph in a more effective way to block the flow of label information from the labeled train nodes. This results in symbiotic attacks being both more robust against larger test sets than plain evasion, while also being stronger than poisoning alone.

4.4 Hyperparameters

Block size. Figure 3 shows the results of the four attacks with varying block sizes, a fixed 5% budget, and 125 attack iterations against a GCN. We observe that for very small block sizes, the attacks are less effective since the PR-BCD optimization can cover only a small part of the adjacency matrix. However, the marginal benefit of larger block sizes decreases once a large part of the adjacency matrix can be covered.

Budget. Considering different fractions of the number of edges as the global budget, Figure 4 indicates that all four attacks follow a similar trend with an increasing budget. Especially on PubMed where the share of labeled train nodes is the lowest, changing 5% of the edges is enough to achieve close to zero accuracy under the symbiotic threat model. This highlights the potentially devastating effect of the joint attacks especially in large graphs with few labeled train nodes.

5 Conclusion / Future Work

In this work, we considered the symbiotic threat model of a combined test and train time adversary aiming to degrade the overall performance of a GNN on node classification tasks. We proposed two methods of obtaining perturbations, one a special case of the other, and demonstrated the potentially devastating impact such an attacker can have compared to a plain poisoning or evasion attacker. We now conclude by outlining potential lines of future work regarding the symbiotic threat model.

Different Attacks as the Inner Evasion Step. The joint attack we described is not limited to PR-BCD, and can be used with other evasion attacks, or attacks designed specifically for the symbiotic setting, to obtain perturbations within the poisoning attack.

Inductive Tasks, Local Budgets, Targeted Attacks. The combined threat model is also applicable to inductive tasks in which the test graph is different from the train graph. The poisoning half of the joint attack can be performed by targeting the validation nodes rather than the test nodes, and the evasion half is performed on the new test graph. Similarly, the attacks can also be applied with small changes under per-node local budgets, or target specific nodes. We plan to include further evaluations on these settings as our next step.

New Evasion-Aware Poisoning Attacks. Finally, further novel poisoning attacks can also be developed which utilize the knowledge of a future evasion in different ways.

Acknowledgments and Disclosure of Funding

This research was supported by the Helmholtz Association under the joint research school "Munich School for Data Science - MUDS".

References

Bojchevski, A. and Günnemann, S. (2019). Adversarial Attacks on Node Embeddings via Graph Poisoning.

Chen, J., Lin, X., Shi, Z., and Liu, Y. (2020). Link prediction adversarial attack via iterative gradient attack. *IEEE Transactions on Computational Social Systems*, 7(4):1081–1094.

Chien, E., Peng, J., Li, P., and Milenkovic, O. (2020). Adaptive Universal Generalized PageRank Graph Neural Network. In *International Conference on Learning Representations*.

Dai, H., Li, H., Tian, T., Huang, X., Wang, L., Zhu, J., and Song, L. (2018). Adversarial Attack on Graph Structured Data. In *Proceedings of the 35th International Conference on Machine Learning*, pages 1115–1124. PMLR.

Gasteiger, J., Bojchevski, A., and Günnemann, S. (2018). Predict then Propagate: Graph Neural Networks meet Personalized PageRank. In *International Conference on Learning Representations*.

Geisler, S., Schmidt, T., Şirin, H., Zügner, D., Bojchevski, A., and Günnemann, S. (2021). Robustness of Graph Neural Networks at Scale. In *Advances in Neural Information Processing Systems*, volume 34, pages 7637–7649. Curran Associates, Inc.

Gosch, L., Geisler, S., Sturm, D., Charpentier, B., Zügner, D., and Günnemann, S. (2023a). Adversarial training for graph neural networks. In *Neural Information Processing Systems, NeurIPS*.

Gosch, L., Sturm, D., Geisler, S., and Günnemann, S. (2023b). Revisiting robustness in graph machine learning. In *11th International Conference on Learning Representations, ICLR*.

Guo, J., Janet, J. P., Bauer, M. R., Nittinger, E., Giblin, K. A., Papadopoulos, K., Voronov, A., Patronov, A., Engkvist, O., and Margreitter, C. (2021). Dockstream: a docking wrapper to enhance de novo molecular design. *Journal of cheminformatics*, 13(1):1–21.

Hao, J., Zhao, T., Li, J., Dong, X. L., Faloutsos, C., Sun, Y., and Wang, W. (2020). P-companion: A principled framework for diversified complementary product recommendation. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2517–2524.

Kipf, T. N. and Welling, M. (2016). Semi-Supervised Classification with Graph Convolutional Networks. In *International Conference on Learning Representations*.

McCallum, A. K., Nigam, K., Rennie, J., and Seymore, K. (2000). Automating the construction of internet portals with machine learning. *Information Retrieval*, 3:127–163.

Mujkanovic, F., Geisler, S., Günnemann, S., and Bojchevski, A. (2022). Are Defenses for Graph Neural Networks Robust? *Advances in Neural Information Processing Systems*, 35:8954–8968.

Nesterov, Yu. (2012). Efficiency of Coordinate Descent Methods on Huge-Scale Optimization Problems. *SIAM Journal on Optimization*, 22(2):341–362.

Sen, P., Namata, G., Bilgic, M., Getoor, L., Galligher, B., and Eliassi-Rad, T. (2008). Collective classification in network data. *AI magazine*, 29(3):93–93.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. (2018). Graph Attention Networks.

Wang, X., Chang, H., Xie, B., Bian, T., Zhou, S., Wang, D., Zhang, Z., and Zhu, W. (2023). Revisiting adversarial attacks on graph neural networks for graph classification. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–12.

Wu, H., Wang, C., Tyshetskiy, Y., Docherty, A., Lu, K., and Zhu, L. (2019). Adversarial Examples on Graph Data: Deep Insights into Attack and Defense.

Xu, K., Chen, H., Liu, S., Chen, P.-Y., Weng, T.-W., Hong, M., and Lin, X. (2019). Topology Attack and Defense for Graph Neural Networks: An Optimization Perspective.

Zhang, H., Zheng, T., Gao, J., Miao, C., Su, L., Li, Y., and Ren, K. (2019). Data poisoning attack against knowledge graph embedding. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, IJCAI'19, pages 4853–4859, Macao, China. AAAI Press.

Zhu, D., Zhang, Z., Cui, P., and Zhu, W. (2019). Robust Graph Convolutional Networks Against Adversarial Attacks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1399–1407, Anchorage AK USA. ACM.

Zügner, D., Akbarnejad, A., and Günnemann, S. (2018). Adversarial Attacks on Neural Networks for Graph Data. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, pages 2847–2856, New York, NY, USA. Association for Computing Machinery.

Zügner, D., Borchert, O., Akbarnejad, A., and Günnemann, S. (2020). Adversarial Attacks on Graph Neural Networks: Perturbations and their Patterns. *ACM Transactions on Knowledge Discovery from Data*, 14(5):57:1–57:31.

Table 3: Perturbed accuracies (± standard error) of the joint and sequential attacks under the symbiotic threat model with a 5% global budget. The -J suffix indicates the graph has been pre-processed with Jaccard purification (Wu et al., 2019).

Model	Dataset	Clean	Sequential	Joint
GCN	CiteSeer	0.68 ± 0.01	0.41 ± 0.01	**0.38 ± 0.01**
	CiteSeer-J	0.68 ± 0.01	0.4 ± 0.01	**0.38 ± 0.01**
	Cora	0.78 ± 0.01	0.37 ± 0.02	**0.35 ± 0.01**
	Cora-J	0.74 ± 0.01	**0.36 ± 0.01**	**0.36 ± 0.02**
	PubMed	0.78 ± 0.01	0.05 ± 0.01	**0.03 ± 0.01**
	PubMed-J	0.77 ± 0.01	0.04 ± 0.01	**0.02 ± 0.0**
GAT	CiteSeer	0.62 ± 0.02	**0.3 ± 0.03**	0.38 ± 0.02
	CiteSeer-J	0.64 ± 0.01	**0.3 ± 0.03**	0.36 ± 0.02
	Cora	0.69 ± 0.02	**0.29 ± 0.02**	0.32 ± 0.02
	Cora-J	0.67 ± 0.01	**0.28 ± 0.02**	0.3 ± 0.03
	PubMed	0.73 ± 0.01	0.24 ± 0.02	**0.2 ± 0.03**
	PubMed-J	0.74 ± 0.01	0.27 ± 0.04	**0.19 ± 0.02**
APPNP	CiteSeer	0.69 ± 0.01	**0.47 ± 0.01**	0.48 ± 0.01
	CiteSeer-J	0.68 ± 0.01	**0.45 ± 0.02**	**0.45 ± 0.02**
	Cora	0.82 ± 0.02	0.54 ± 0.02	**0.51 ± 0.04**
	Cora-J	0.82 ± 0.01	0.57 ± 0.01	**0.54 ± 0.01**
	PubMed	0.79 ± 0.0	**0.09 ± 0.02**	**0.09 ± 0.01**
	PubMed-J	0.77 ± 0.01	**0.1 ± 0.02**	0.12 ± 0.02
GPRGNN	CiteSeer	0.66 ± 0.01	0.34 ± 0.01	**0.33 ± 0.01**
	CiteSeer-J	0.65 ± 0.01	**0.35 ± 0.01**	**0.35 ± 0.01**
	Cora	0.82 ± 0.01	0.41 ± 0.01	**0.4 ± 0.01**
	Cora-J	0.79 ± 0.01	0.42 ± 0.01	**0.4 ± 0.01**
	PubMed	0.78 ± 0.01	**0.08 ± 0.02**	0.11 ± 0.03
	PubMed-J	0.78 ± 0.01	0.16 ± 0.05	**0.15 ± 0.04**
RGCN	CiteSeer	0.63 ± 0.01	**0.47 ± 0.01**	**0.47 ± 0.01**
	Cora	0.74 ± 0.02	0.56 ± 0.01	**0.52 ± 0.02**
	PubMed	0.77 ± 0.01	0.28 ± 0.04	**0.15 ± 0.03**

Figure 5: Perturbed accuracies and standard error after the joint attack on a GCN, with the x-axis indicating the number of iterations for the inner evasion part of the joint attack.

A Comparing the Joint and Sequential Attacks

Extending on Table 2, Table 3 further compares the joint and sequential attacks of the symbiotic threat model. Although the joint attack is stronger than the sequential attack in a higher number of cases, they are often within one standard error of each other so it is difficult to argue for a statistically significant difference. This might indicate the difficulty of estimating the future evasion perturbations

Table 4: Inductive

Model	Dataset	Clean	Evasion	Poisoning	Symbiotic
GCN	CiteSeer	0.67 ± 0.01	0.41 ± 0.01	0.62 ± 0.01	**0.33 ± 0.01**
	Cora	0.75 ± 0.02	0.42 ± 0.01	0.68 ± 0.03	**0.3 ± 0.01**
GAT	CiteSeer	0.68 ± 0.01	**0.37 ± 0.01**	0.64 ± 0.02	0.56 ± 0.02
	Cora	0.77 ± 0.01	**0.21 ± 0.01**	0.61 ± 0.04	0.35 ± 0.03
APPNP	CiteSeer	0.71 ± 0.01	0.47 ± 0.01	0.66 ± 0.02	**0.4 ± 0.01**
	Cora	0.82 ± 0.02	0.53 ± 0.02	0.78 ± 0.01	**0.37 ± 0.01**
GPRGNN	CiteSeer	0.67 ± 0.01	0.37 ± 0.01	0.56 ± 0.01	**0.34 ± 0.01**
	Cora	0.8 ± 0.02	0.44 ± 0.01	0.74 ± 0.01	**0.35 ± 0.01**

from within the poisoning attack due to the very large search space, and highlight the space for potential developments of stronger attacks.

B Inner Evasion Iterations.

The number of iterations the evasion attack runs for within the joint attack is also configurable. Figure 5 displays the accuracy of a GCN on our three benchmark graphs after the joint attack with different number of inner-evasion iterations. We observe, contrary to our expectations, that increasing the number of iterations for the inner-evasion attack has no significant influence on the perturbed accuracy. This perhaps highlights a potential point of improvement for future symbiotic attacks.