– For **Reviewer # 1**: Thank you for the comments! We address your specific concerns in detail below. –

Q1. Fully black-box. We first note that the mild use of model information to identify features to perturb can be replaced with domain knowledge in real applications (line 287-294), which will make the black-box setup strict. To further verify this, we construct a synthetic dataset where we know which features are important for the labels. GC-RWCS works better than all other baselines on this synthetic data without any access to the model information. The full results cannot be shown here due to the page limit, but we will include them in the Appendix of the final draft.

Q2. The constant perturbation vector. The perturbation vector indeed could be optimized given more domain knowledge. We use a constant vector to reflect "very limited attacker knowledge" and leave optimizing it as future work.

Q3. The maximum degree. The maximum degree m does not influence the single node effect (proposition 1) but surely has an impact on the overall attack performance. This impact is hard to quantify theoretically, which is an interesting future direction. If one were able to quantify it, the effect of m should be reflected in the greedy correction steps.

– For **Reviewer # 2**: Thank you for the comments! We address your specific concerns in detail below. –

Q1. More GNNs. Per your advice, we conduct the same experiments on the Graph Attention Network (GAT). We observe a similar trend: the proposed GC-RWCS strategy is able to significantly degrade the model performance; and it outperforms all the baseline methods (see the table below). We will incorporate these experiments in our final draft.

Dataset: threshold	Cora: 10%	20%	30%	Citeseer: 10%	20%	30%	Pubmed: 10%	20%	30%
No-Attack		87.8 ± 0.2			76.9 ± 0.3			85.2 ± 0.1	
Random	72.9 ± 0.5	73.8 ± 0.6	73.9 ± 0.6	70.0 ± 0.5	71.2 ± 0.4	71.7 ± 0.4	73.9 ± 0.4	75.4 ± 0.3	76.2 ± 0.3
Degree	66.5 ± 0.7	67.3 ± 0.7	69.8 ± 0.7	63.3 ± 0.5	65.9 ± 0.4	67.9 ± 0.3	66.7 ± 0.7	69.0 ± 0.5	71.2 ± 0.4
Pagerank	74.3 ± 0.5	74.8 ± 0.3	82.4 ± 0.2	69.5 ± 0.3	72.9 ± 0.3	74.2 ± 0.3	71.6 ± 0.4	78.1 ± 0.2	79.1 ± 0.2
Betweenness	64.8 ± 0.5	66.0 ± 0.5	67.3 ± 0.6	65.2 ± 0.5	66.5 ± 0.4	67.6 ± 0.3	63.4 ± 0.7	68.4 ± 0.6	72.0 ± 0.4
RWCS	71.1 ± 0.5	74.6 ± 0.3	82.5 ± 0.2	69.2 ± 0.3	72.9 ± 0.3	73.9 ± 0.3	69.4 ± 0.5	74.9 ± 0.3	77.9 ± 0.2
GC-RWCS	**58.1** ± 0.6*	**57.9** ± 0.6*	**63.0** ± 0.5*	**58.3** ± 0.6*	**61.9** ± 0.6*	**61.9** ± 0.4*	**58.9** ± 0.9*	**63.8** ± 0.7*	**68.9** ± 0.5*

Q2. Figure 1, and RWCS vs PageRank. First, we clarify that PageRank is not missing but almost overlaps with RWCS in Figure 1. RWCS is indeed very similar to PageRank (line 188). However, we also highlighted our contribution (line 57-59) for **revealing the novel connection** between the black-box adversarial attack on GNN and the PageRank-like heuristic, RWCS. Thanks to this connection, we further developed the practically effective strategy, GC-RWCS.

Q3. Sensitivity of parameter k. We first note that k **was fixed as 1 in all experiment setups** in our paper (see line 276). Responsively, we further conduct a sensitivity analysis of k. We observe similar trends on all datasets and thresholds, and below we show results on Cora with threshold 30% due to the page limit. The results under GC-RWCS attacks with $k = 1$ and $k = 2$ are very similar. The results of the null choice $k = 0$, i.e., not removing neighbors, are slightly worse as expected; but they are still better than all other baselines.

Model	GCN			GAT			JKNetConcat		
k	$k=0$	$k=1$	$k=2$	$k=0$	$k=1$	$k=2$	$k=0$	$k=1$	$k=2$
GC-RWCS	81.0 ± 0.5	**80.7** ± 0.5	**80.7** ± 0.5	66.1 ± 0.7	63.0 ± 0.5	**62.6** ± 0.8	65.2 ± 1.0	**59.1** ± 1.6	62.1 ± 1.3

– For **Reviewer # 3**: Thank you for the comments! We address your specific concerns in detail below. –

Q1. Assumption 5 is strong. We agree that assumption 5 (which comes from [22]) seems a bit strong. However, we believe assumption 5 approximately holds **at a coarse level**, which is enough to develop a **black-box** attack strategy. **And our empirical results indeed seem to support our conjecture.**

Q2. Influence of attack set constraints. We would like to clarify that the constraints on the attack set are illustrated by the optimization problem (2) after line 145, which are $|S| \leq r, d_i \leq m, \forall i \in S$. **The influence of these constraints on the proposed method is fundamental**: these constraints define the optimization problem (2) and its black-box counterpart (a novel black-box attack setup), which then lead to the derivation of the proposed method.

Q3. Description of baselines. The three baselines are strategies that "select nodes with top centrality" (line 286), with the centrality metrics being Degree, Betweeness, and PageRank. We will add more details in final draft.

Q4. Comparison to [1, 3]. [1] and [3] require extra model information thus not applicable in our setup (line 266-268).

Q5. Minor improvements over baselines. We respectfully disagree. GC-RWCS performs the best **in all but one setup**. The difference between GC-RWCS and the best baseline is **significant in 15/18 setups**, with **up to 13%** improvement.

Q6. J as attack strength. We verified that replacing λ with J as the measure for attack strength yields almost the same plots in Figure 1. We are not able to include the new plots here due to the page limit but will do so in Appendix.

Q7. Why attack is less effective on GCN than on JKNets. Our intuition is that GCN is shallower than JKNets so less model information is leaked by the graph structure. This phenomenon also supports our claim in line 68-69.

– For **Reviewer # 4**: Thank you for your suggestion! Adversarial defense under our novel black-box setup is indeed an interesting future direction. For optimizing perturbation matrix please see our response to Q2 of Reviewer 1.