Supplement

Formulation of Convex Quadratic Program

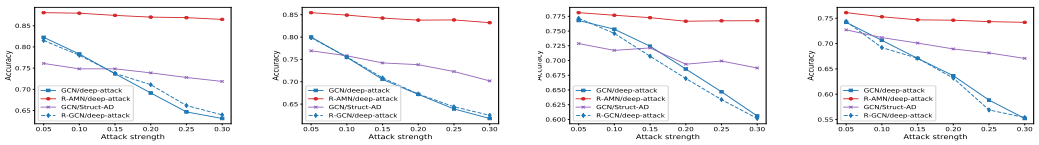
We explicitly write out the convex quadratic program for learning robsut AMN, which is omitted in the main paper. By LP duality, we can replace the attacker's maximization problem using its dual minimization problem, which is further integrated into Eqn. (5). Consequently, we can approximate Eqn. (5) by the following convex quadratic program:

$$
\begin{aligned}
&\min \ \frac{1}{2}||\mathbf{w}||^2 + C(N - \sum_{i=1}^{N}\sum_{k=1}^{K} \mathbf{w}_n^k \mathbf{x}_i \hat{y}_i^k + \sum_{i=1}^{N} t_i + \sum_{(i,j)\in E} p_{ij} - D^- \cdot t_D) \\
&\text{s.t.} \quad \forall i,k, \quad t_i - \sum_{(i,j),(j,i)\in E} t_{ij}^k - \mathbf{w}_n^k \mathbf{x}_i + \hat{y}_i^k \geq 0, \\
&\forall (i,j) \in E, k, \quad s_{ij}^k + t_{ij}^k + t_{ji}^k - w_e^k \geq 0, \ s_{ij}^k, t_{ij}^k, t_{ji}^k \geq 0, \\
&\forall (i,j) \in E, \quad p_{ij} - \sum_{k=1}^{K} s_{ij}^k - t_D + \sum_{(i,j)\in E}\sum_{k=1}^{K} w_e^k \hat{y}_i^k \hat{y}_j^k \geq 0, \ p_{ij}, t_D \geq 0.
\end{aligned} \tag{1}
$$

The minimization is over the weights $\mathbf{w}$ and the dual variables $t_i, p_{ij}, s_{ij}^k, t_{ij}^k, t_{ji}^k, t_D$.

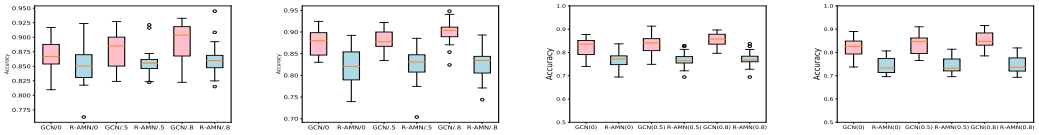
Additional Experiment Results

We compare R-AMN and GCN under the deep-attack as well as on non-adversarial data on the Cora and CiteSeer datasets in the same experiment settings as in the main paper. Specifically, in Fig. 1, "R-AMN/deep-attack" shows the accuracies of R-AMN under deep-attack with various degrees of graph perturbations, where the train graph and test graph are attacked by deep-attack separately. It demonstrates that R-AMN is robust to deep-attack even with relatively large structural perturbations. "GCN/deep-attack" and "GCN/Struct-AD" show the accuracies of GCN under deep-attack and our proposed *Struct-AD* attack, respectively. Generally, deep-attack is a much more effective method to attack GCN models. Fig. 2 demonstrated that on non-adversarial data, the performances of R-AMN and GCN are comparable.



(a) Cora-H (b) Cora-L (c) CiteSeer-H (d) CiteSeer-L

Figure 1: R-AMN and R-GCN under deep attack; GCN under deep attack and Struct-AD attack.



(a) Cora-H (b) Cora-L (c) CiteSeer-H (d) CiteSeer-L

Figure 2: R-AMN and GCN on non-adversary data as graphs are purified, e.g. R-AMN(0.5) stands for R-AMN when noisy edges are deleted with probability 0.5.