Table 1: Mismatching rate (%) with 5% perturbed edges

	AS			**SNS**			**DBLP**		
Attack Model	**SNNA**	**CrossMNA**	**DGMC**	**SNNA**	**CrossMNA**	**DGMC**	**SNNA**	**CrossMNA**	**DGMC**
Clean	53.9	46.6	34.7	45.2	50.4	41.6	56.1	51.9	63.2
GMA+Robust Training [7]	62.6	58.5	53.0	56.2	66.5	51.8	71.8	71.0	77.3
Vaccinated GMA [8]	62.1	59.9	53.2	58.7	67.6	54.6	70.4	68.6	74.9
GMA	**64.2**	**62.9**	**54.9**	**61.2**	**69.6**	**55.7**	**74.2**	**74.3**	**80.7**

We would like to thank the four reviewers for the helpful and constructive comments. We have tried our best to clarify the concerns and comments by all four reviewers.

1. Ablation study of unnoticeable perturbations and KDE (Reviewers: 1+3)

We have included the ablation study in Figure 5 in Page 8 in the submission. Compared with our proposed GMA model with the full support of both KDE and MLPGD components, one variant, GMA-KDE, only uses our proposed KDE and density maximization to generate imperceptible attacks. GMA-KDE achieves the better attack performance than GMA-MLPGD, another version that only employs our proposed MLPGD to well choose good attack starting points. A rational guess is that it is difficult to correctly match two nodes when they lie in dense regions with many similar neighbors, although the main goal of KDE is to generate imperceptible attacks.

2. Gaussian model on parameter estimation (Reviewers: 1+4)

Many papers assume graph representations follow Gaussian distributions, which allows to capture graph dynamics and uncertainty [1-4]. We have run the Shapiro-Wilk test, where if the P-Value of test > 0.05, the data is thought to follow a Gaussian distribution. In our test, the P-Values on AS and SNS are 0.668 and 0.543, which indicate two graphs follow Gaussian distribution. We will include the results in the submission. [1] Variational Graph Auto-Encoders, NIPS 2016. [2] Adversarial Network Embedding. AAAI 2018. [3] Deep Gaussian Embedding of Graphs: Unsupervised Inductive Learning via Ranking. ICLR 2018. [4] Dynamic Embedding on Textual Networks via a Gaussian Process. AAAI 2020.

3. Time complexity of the attack method (Reviewer 2)

Based on [6], the complexity of meta learning is $O(d^2)$, where d is the problem dimension. In our case, it is the number of nodes in each graph (N). Both density estimation and PGD have complexity of $O(N^2)$. Thus, the overall complexity is $O(N^2)$, which is the same as most existing attack methods that search the entire graphs to find the weakest edges to attack. [6] On the Convergence Theory of Gradient-Based Model-Agnostic Meta-Learning Algorithms. AISTATS 2020.

4. Validate attack performance under defenses (Reviewer 2)

We test two recent defense methods on generated attacks by our GMA model: [7] uses min-max adversarial training for defense and [8] vaccinates attack with low-rank approximations. As shown in Table 1, even with the defense, our GMA model can still achieve very high mismatching rate. We will include these results in the submission. [7] Topology Attack and Defense for Graph Neural Networks: An Optimization Perspective, IJCAI 2019. [8] All You Need is Low (Rank): Defending Against Adversarial Attacks on Graphs, WSDM 2020.

5. Difference between our model and [5] that leverages meta-learning to generate attacks (Reviewer 2)

[5] conducts adversarial attacks on global node classification of a single graph. It aims to solve a bilevel optimization problem: (1) training classification on graphs and (2) attacking graphs. It gradually improves attack performance by using meta learning to iteratively solve the above two problems. Our model utilizes meta learning to find good attack starting points in two graphs. [5] Adversarial Attacks on Graph Neural Networks via Meta Learning, ICLR 2019.

6. The motivation and transferability of imperceptible attacks generated by KDE (Reviewer 3)

To our best knowledge, our work is the first to integrate small attack budget and density estimation and maximization to produce imperceptible attacks. Real-world graphs often have imbalanced node degree distribution, i.e., some nodes have low degree but some have high degree. Only considering the budget is not enough. For example, removing 3 edges from a high-degree node with 20 edges is imperceptible. However, for a low-degree node with only 2 edges, even if removing only 1 edge, the change is obvious. On the other hand, from the viewpoint of density, if 100 people join a square with 10K people, the change is unnoticeable. But even if 1 person enters an empty square, the change is obvious. The above motivation is based on the intrinsic characteristics of real graphs and is irrelevant to any graph learning tasks. Thus, it is naturally transferred to other graph applications. Moreover, we have evaluated the attack performance of GMA on two applications of node classification network embedding in pp. 5-6 in the supplementary materials.

7. Baseline selection in the experiments (Reviewer 3)

The majority of existing efforts focus on adversarial attacks on single graph learning, especially node classification. To our best knowledge, this work is the first to study adversarial attacks on graph matching. There are no other graph matching baselines available. In fact, we have replaced the original losses in the baselines with the matching loss for fair comparison in our current experiments in the submission.