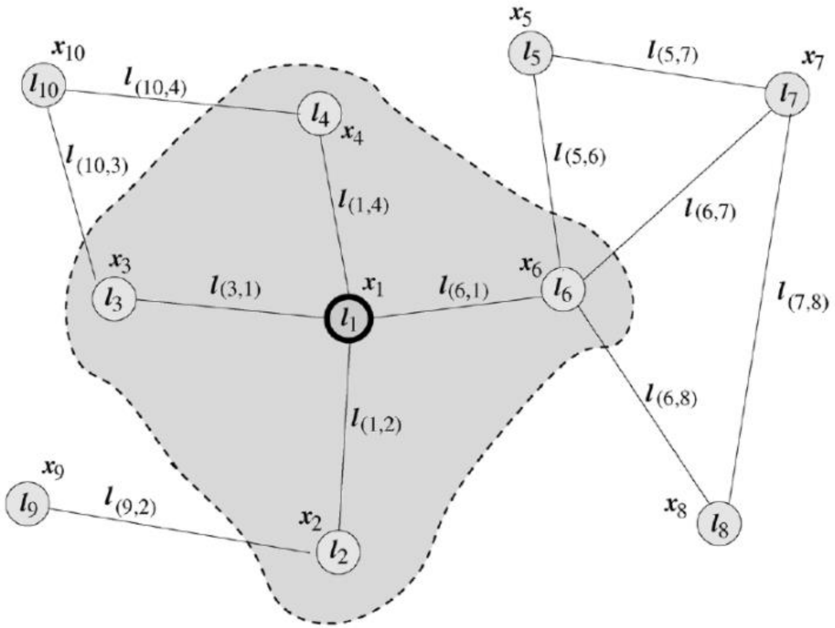
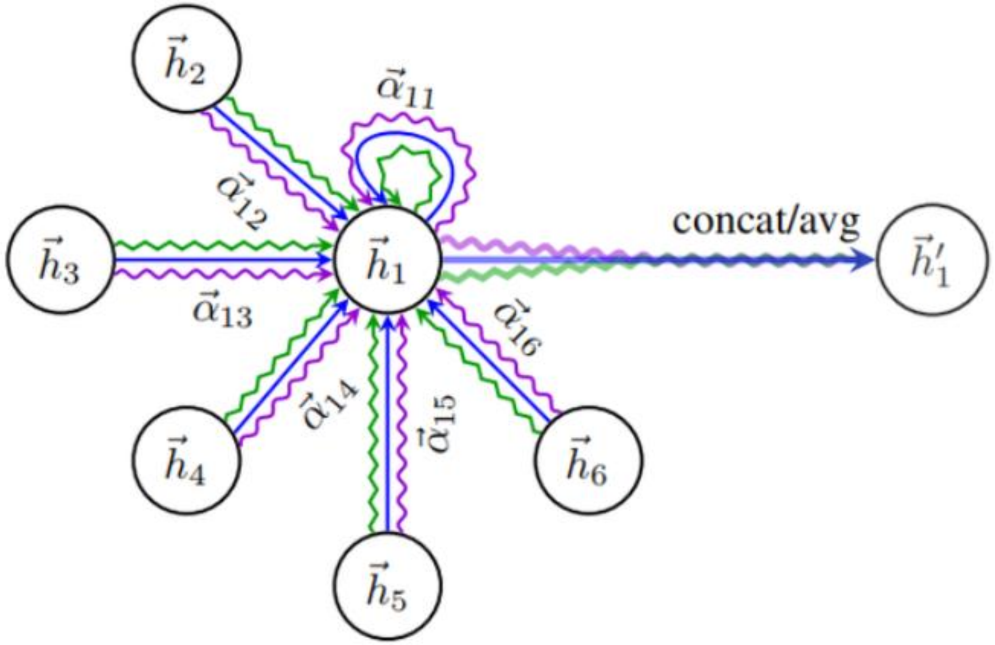
UAG: Uncertainty-aware Attention Graph Neural Network for Defending Adversarial Attacks

Boyuan Feng, Yuke Wang, Yufei Ding

AAAI 2021

Motivation





GNN's robustness is worried about under the critical settings

The main reason is that existing GNNs usually do not provide the uncertainty on the predictions.

Framework

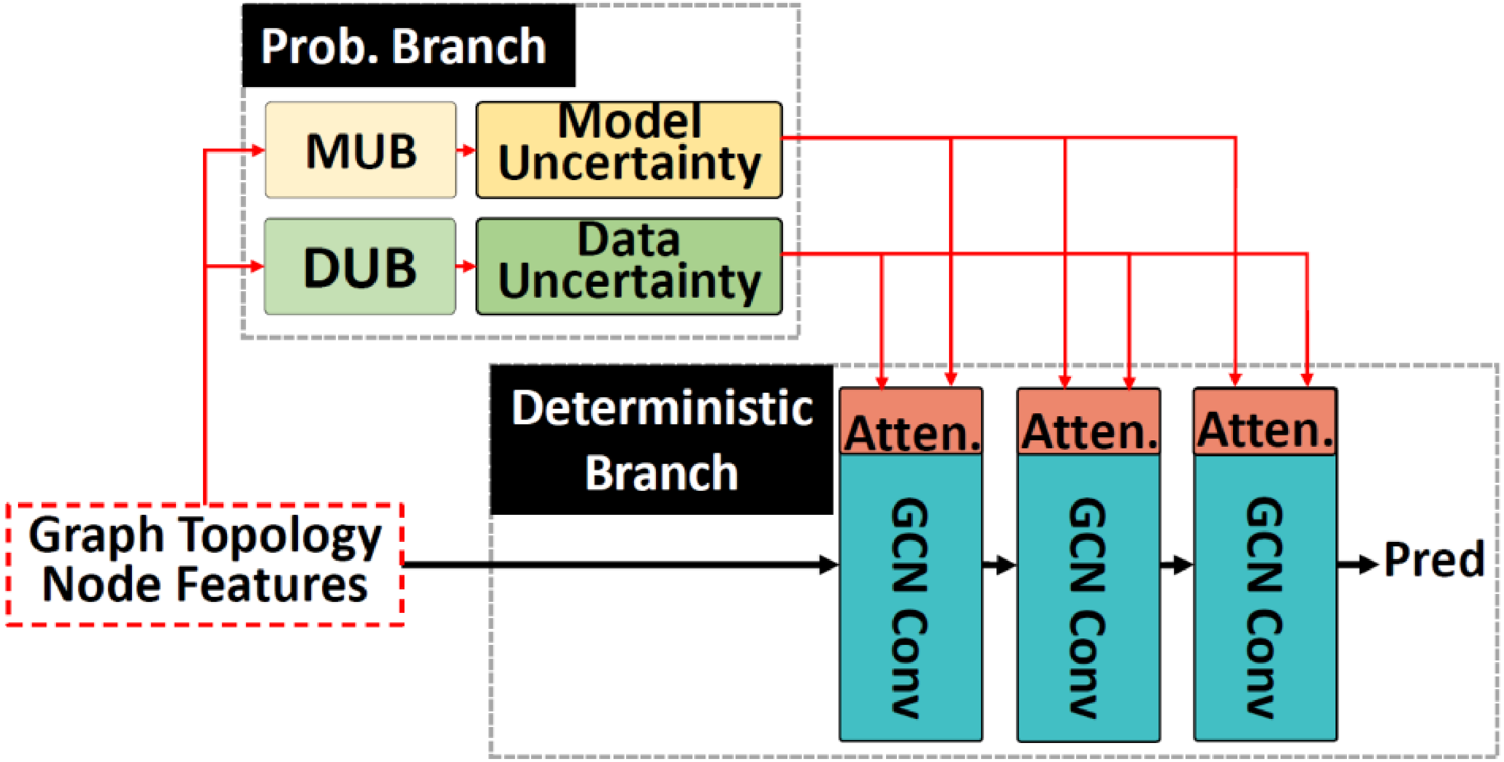


Figure 1: Overview of UAG.

Model uncertainty:

train a two-layer GCN model

model parameter

$$q(W) \sim B \odot W_{MUB}$$
$$P(B) \sim Bernoulli(p) \quad (4)$$

$$L_{model} = -\frac{1}{T}\sum_{t=1}^{T} log\, p(\hat{Y}_t|\hat{W}_t, A, X) + \frac{1-p}{2T}||W_{MUB}||^2 \quad (5)$$

ICLR 2016: Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning prediction result:

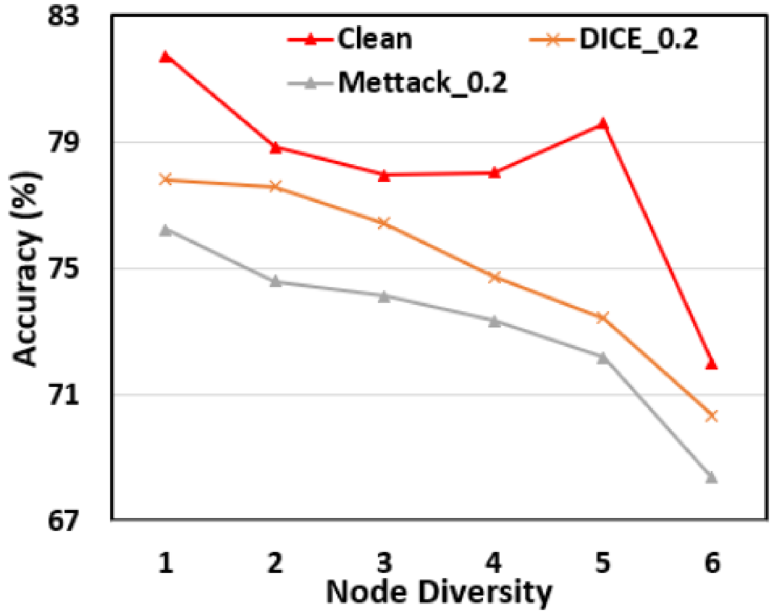
$$E(Y|A, X) = \frac{1}{T} \sum_{t=1}^{T} \hat{Y}_t \tag{6}$$

uncertainty value:

$$\begin{aligned} U_M(Y|A, X) &= Var(Y|A, X) \\ &= E(Y^2|A, X) - [E(Y|A, X)]^2 \\ &= \frac{1}{T} \sum_{t=1}^{T} \hat{Y}_t^2 - [E(Y|A, X)]^2 \end{aligned} \tag{9}$$

Data Uncertainty

Node diversity is data uncertainty's value



label

treating the prediction as a Gaussian distribution and setting the variance to be the node diversity

$$Y \sim N(\hat{\mu}(A, X), \hat{\sigma}^2(A, X)) \quad (10)$$

uncertainty value: $$U_D(Y|A, X) = \hat{\sigma}^2(A, X) \tag{11}$$

labeled data loss: $$L_1 = KL(N(\hat{\mu}(A, X), \hat{\sigma}^2(A, X))|N(Y, \sigma^2)) \tag{12}$$

unlabeled data loss: $$L_2 = \sum_i \sum_{k<l} \sum_{j_k \in N_{ik}} \sum_{j_l \in N_{il}} (E_{ij_k}^2 + exp^{-E_{ij_l}}) \tag{13}$$

$$E_{ij} = D_{KL}(N(\hat{Y}_j, \hat{\sigma}_j^2)||N(\hat{Y}_i, \hat{\sigma}_i^2))$$

attribution value:

$$Att_\tau(u) = exp(-\zeta \cdot U_{\tau,u})$$
$$\zeta = \alpha_\tau \cdot exp(-\beta_\tau \cdot Range(U_\tau)) \qquad (15)$$

$$Att_{Both}(u) = min(Att_M, Att_D) \qquad (16)$$

GNN layer:

$$h_v^{(k+1)} = \sigma(\sum_{u \in \bar{N}(v)} Att_\tau^{uv} \cdot h_u^{(k)} \cdot W^{(k)})$$
$$Att_\tau^{uv} = min(Att_\tau(u), Att_\tau(v)) \qquad (14)$$

Experiments

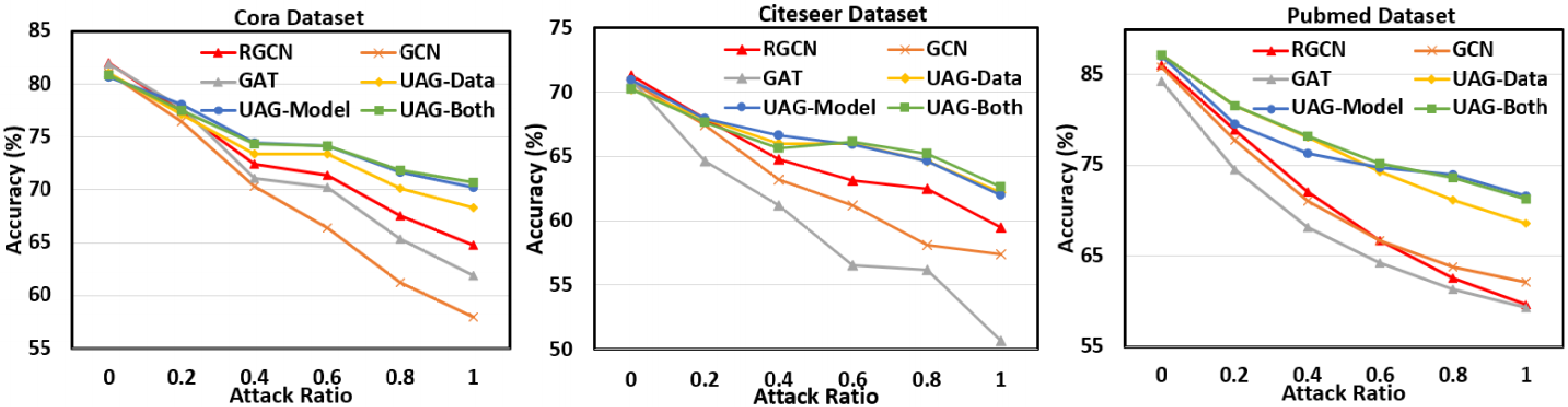


Figure 5: Results of different methods when adopting Random Attack as the attack method.

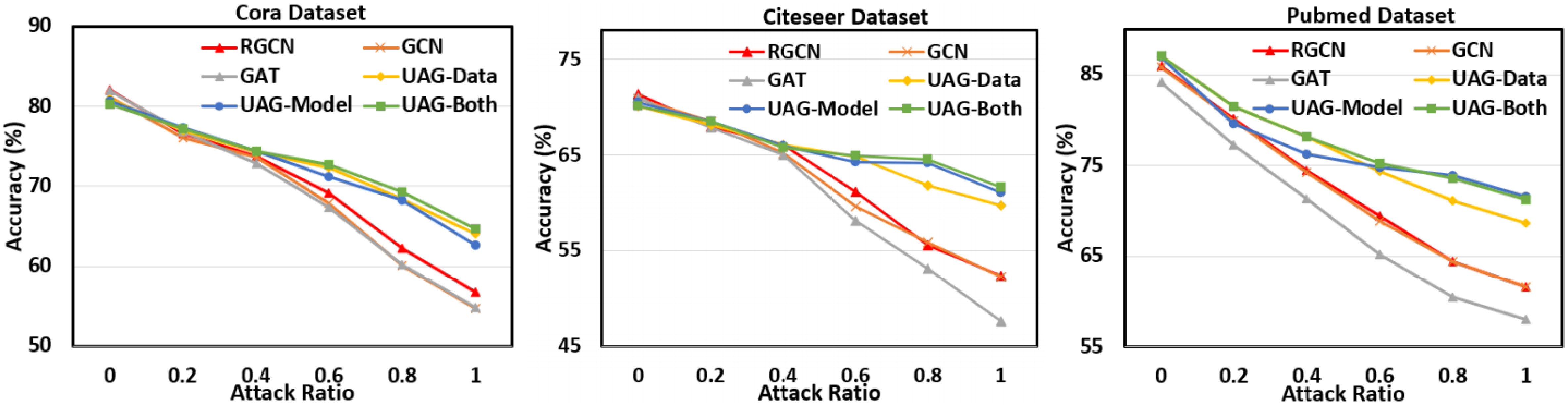


Figure 6: Results of different methods when adopting DICE Attack as the attack method

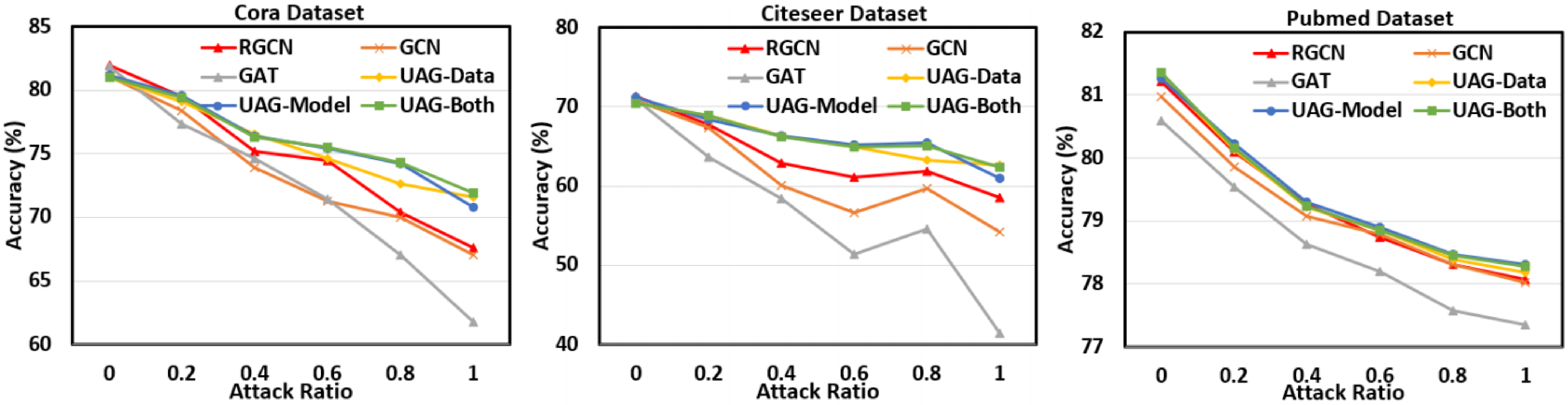


Figure 7: Results of different methods when adopting Mettack as the attack method.

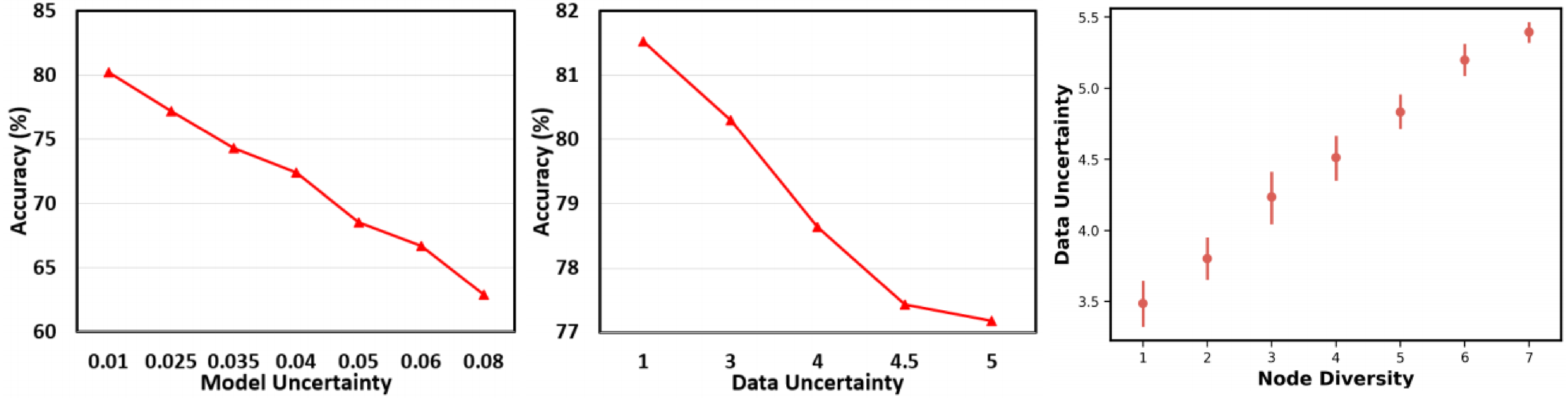
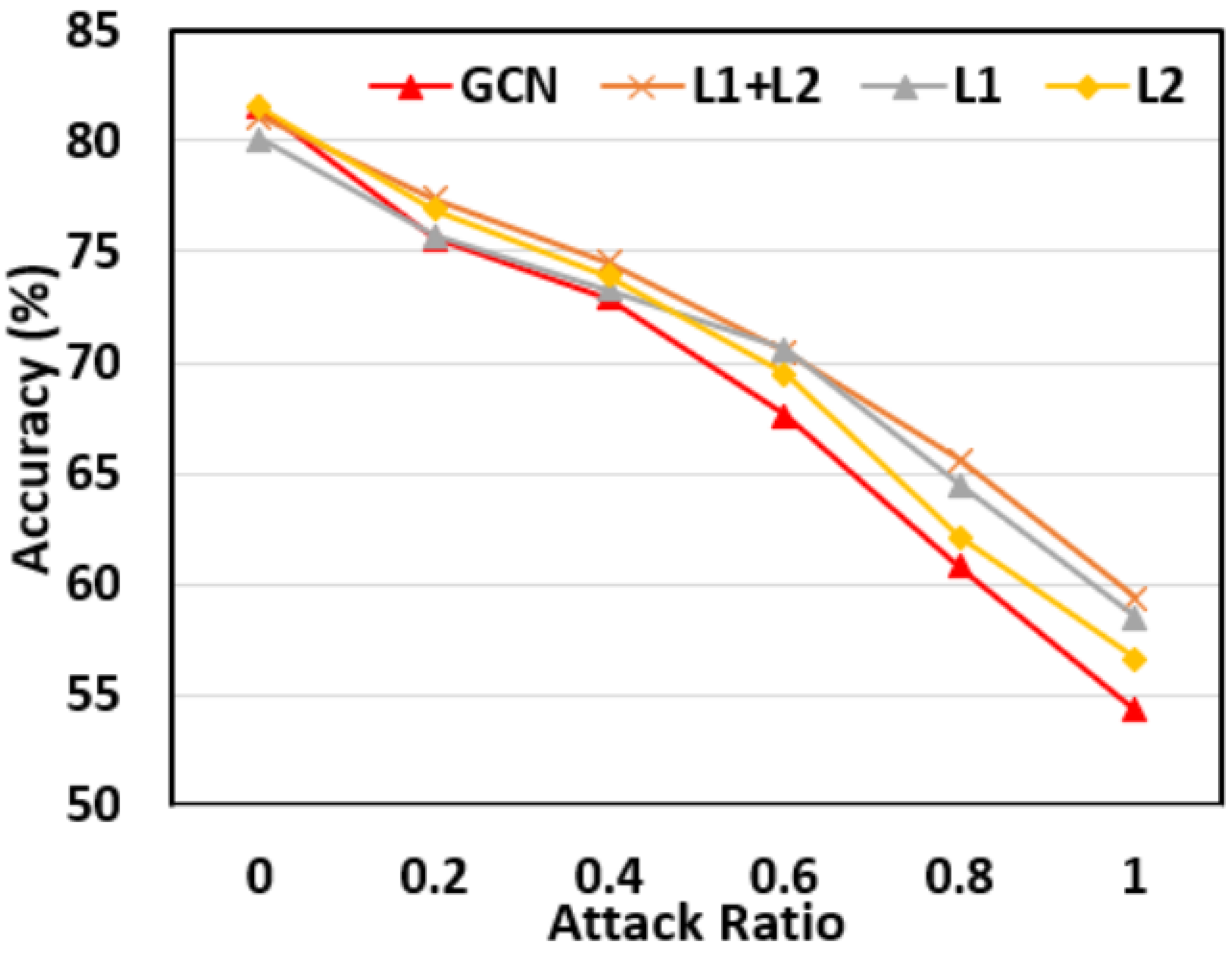


Figure 8: Relationship between Accuracy and Uncertainty. Left: Model Uncertainty v.s. Accuracy. Mid: Data Uncertainty v.s. Accuracy. Right: Data Uncertainty v.s. True Diversity.



(a) Benefit of Loss Designs.

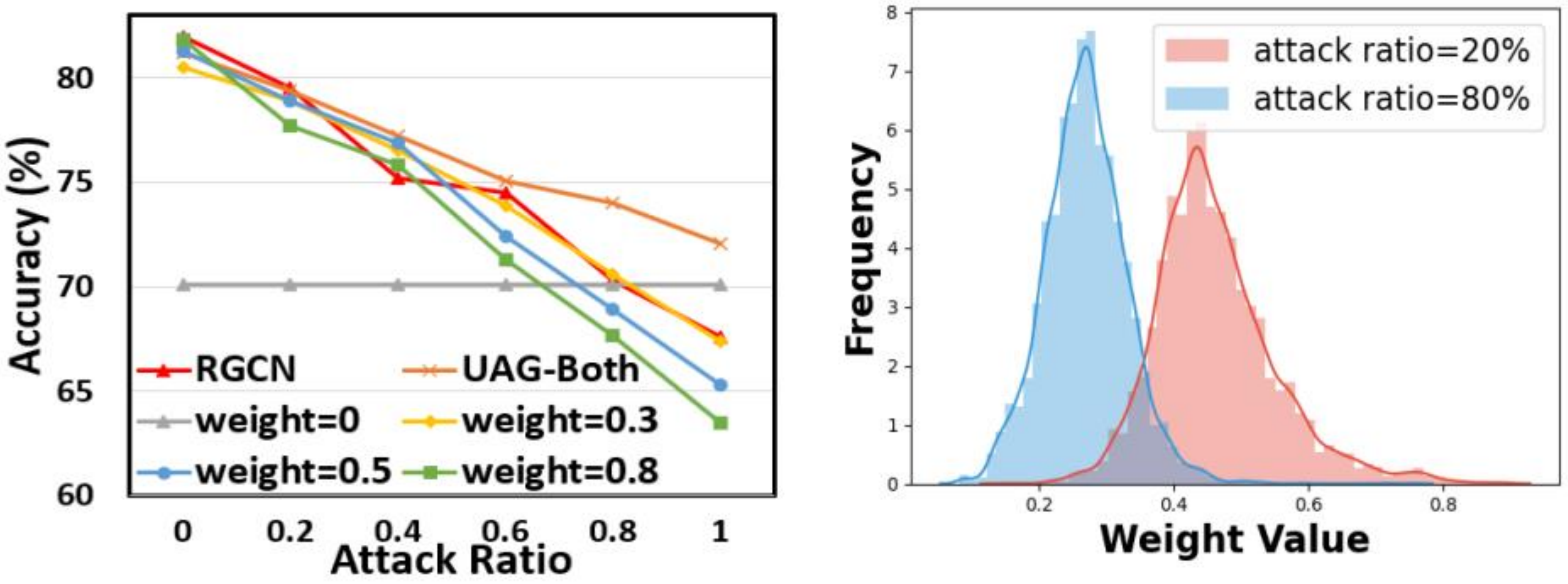


Figure 10: Left: Accuracy of Static Edge Weights. Right: Edge weight distribution under Random Attack.