Robust Heterogeneous Graph Neural Networks against Adversarial Attacks

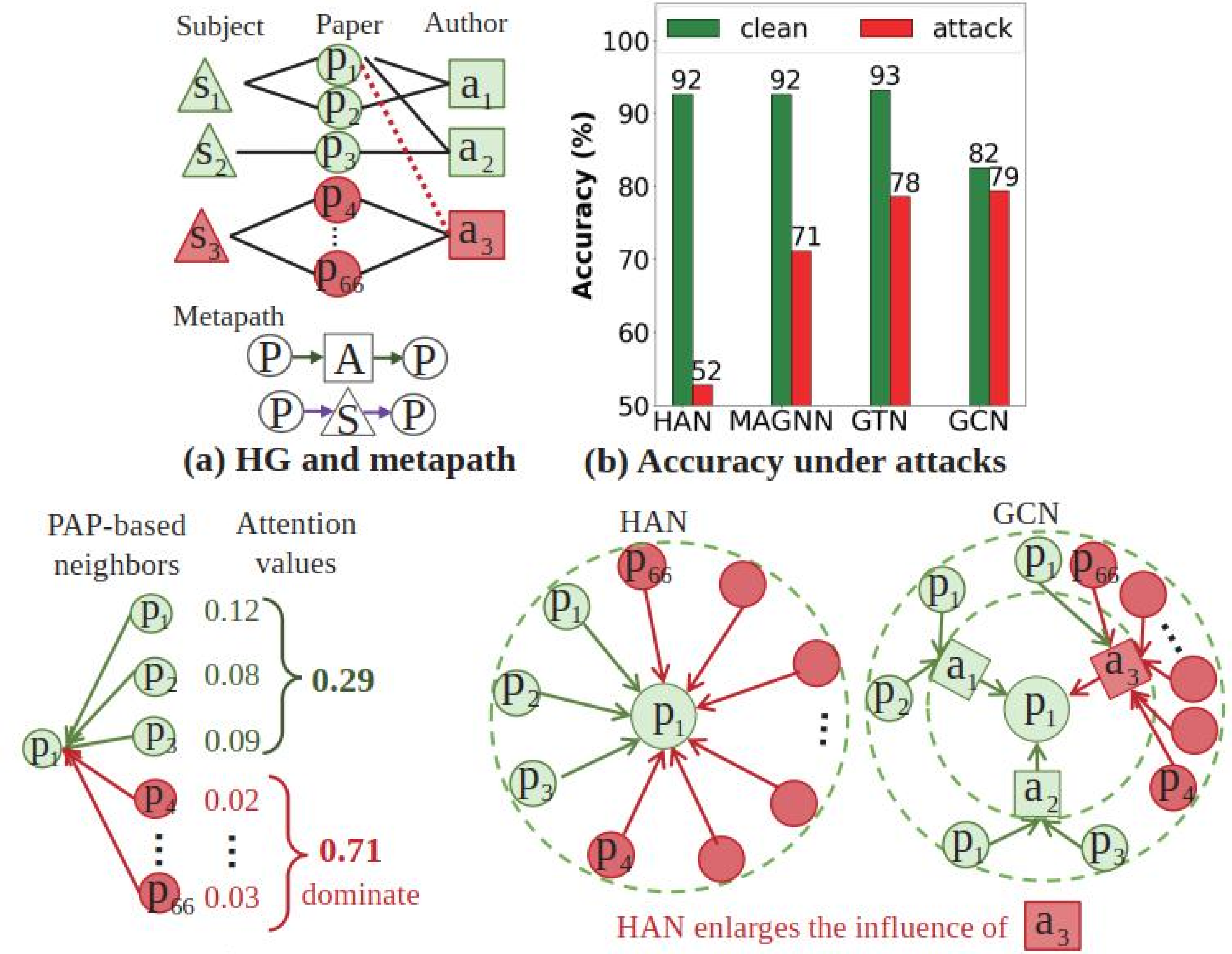
Mengmei Zhang, Xiao Wang, Meiqi Zhu, Chuan Shi, Zhiqiang Zhang, Jun Zhou

Beijing University of Posts and Telecommunications
Ant Group, Hangzhou, China

AAAI 2022

Zhuomin Chen
2022.07.10

Introduction



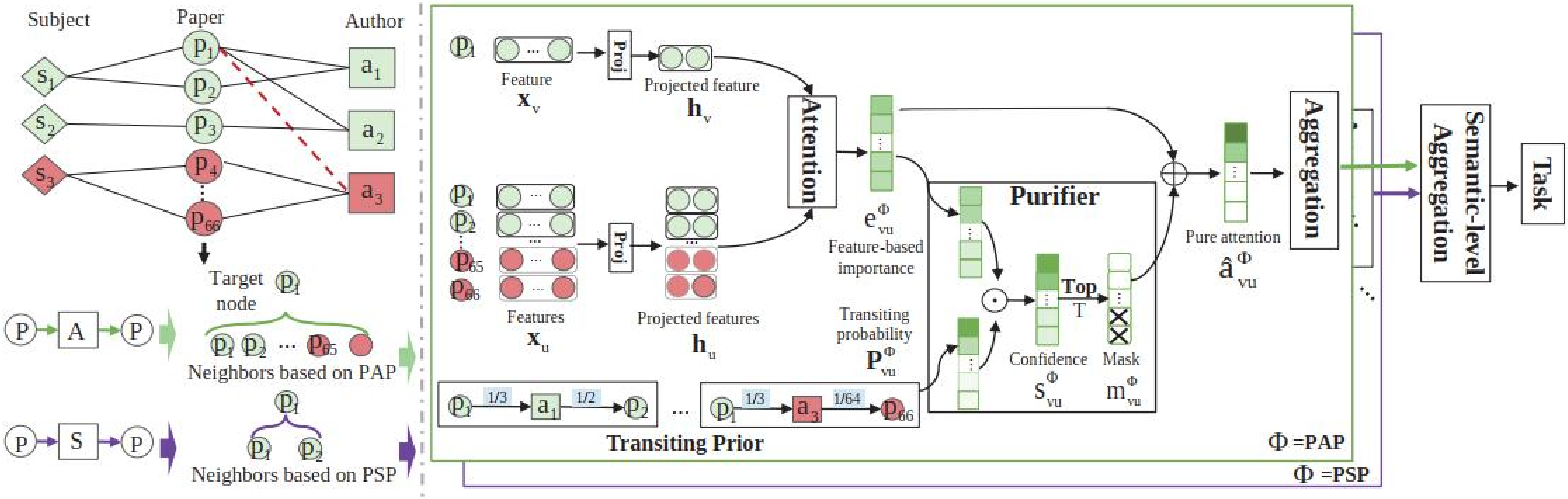
(a) HG and metapath

(b) Accuracy under attacks

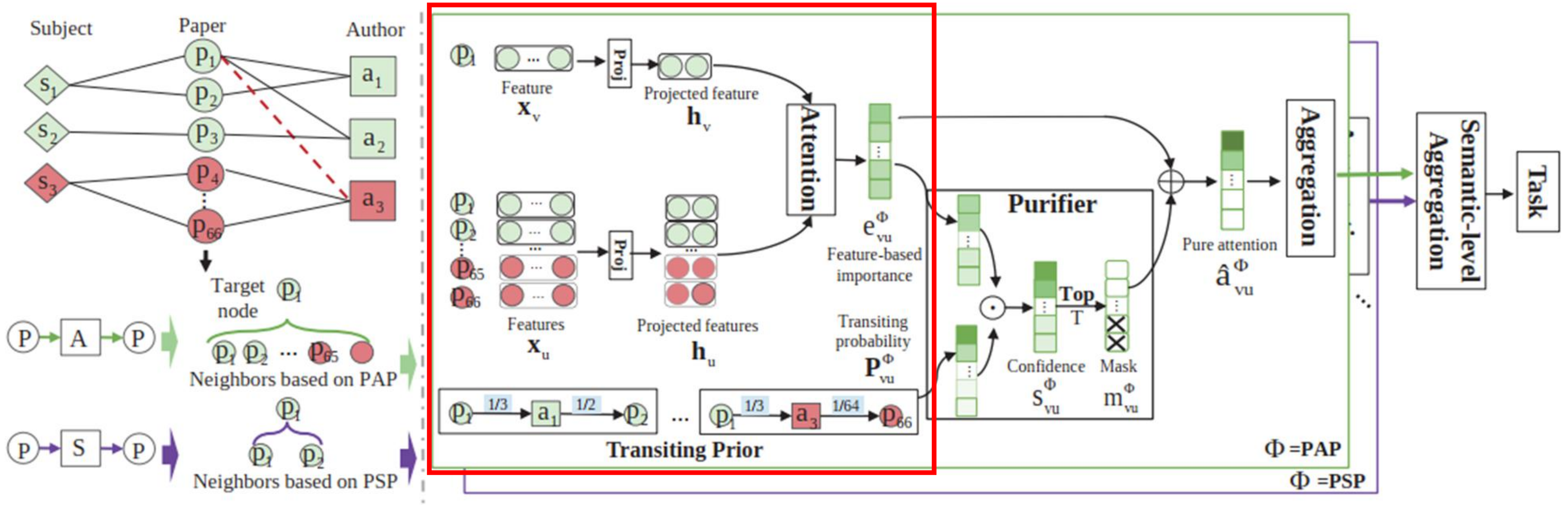
(c) Soft attention values of HAN

(d) Perturbation enlargement effect

Methodology



Node-level Aggregation



Node feature transformation. $\mathbf{h}_v = \mathbf{W}_A \mathbf{x}_v.$ (4)

Feature-based importance. $e^{\Phi}_{vu} = \mathbf{h}_v \cdot \mathbf{h}_u,$ (5)

neighbor $u \in \mathcal{N}^{\Phi}_v$

the importance e^{Φ}_{vu} of neighbors u to target node v under Φ

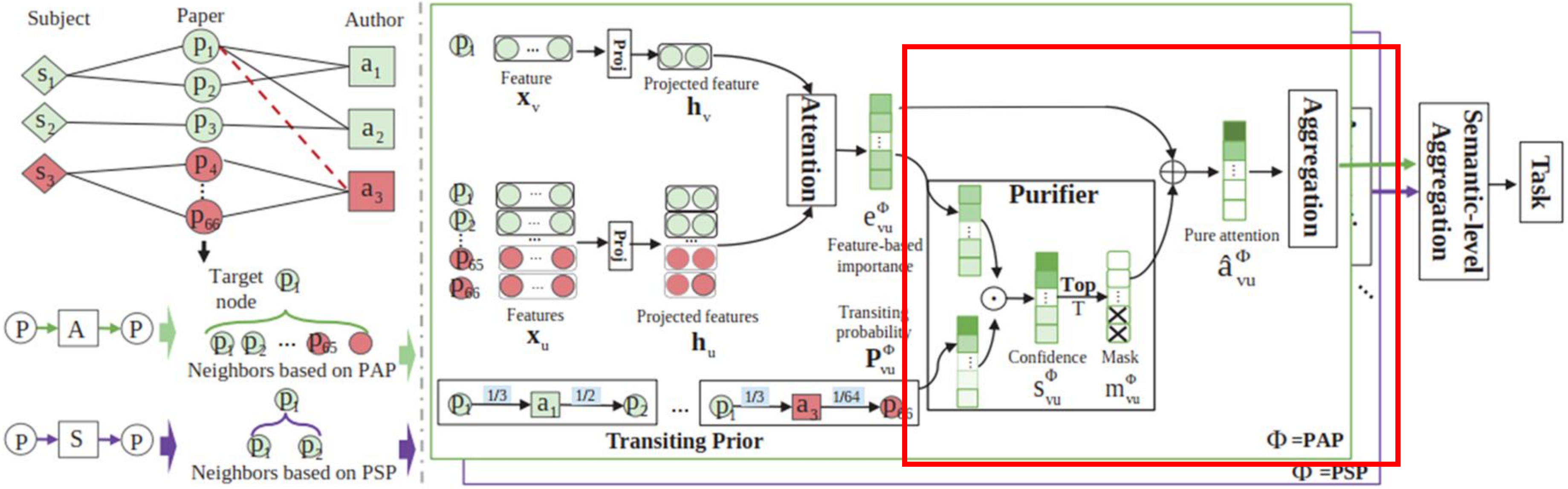
A metapath Φ

$\Phi = A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \cdots \xrightarrow{R_l} A_{l+1}$

Transiting prior. $\mathbf{P}^{\Phi} = \mathbf{P}^{R_1} \cdots \mathbf{P}^{R_l},$ (1)

$\mathbf{P}^{R_i} = (\mathbf{D}^{R_i})^{-1}\mathbf{M}^{R_i}$ Each element $\mathbf{P}^{R_i}_{vu}$ represents the prob-ability of transiting from node v to u in relation R_i

Node-level Aggregation



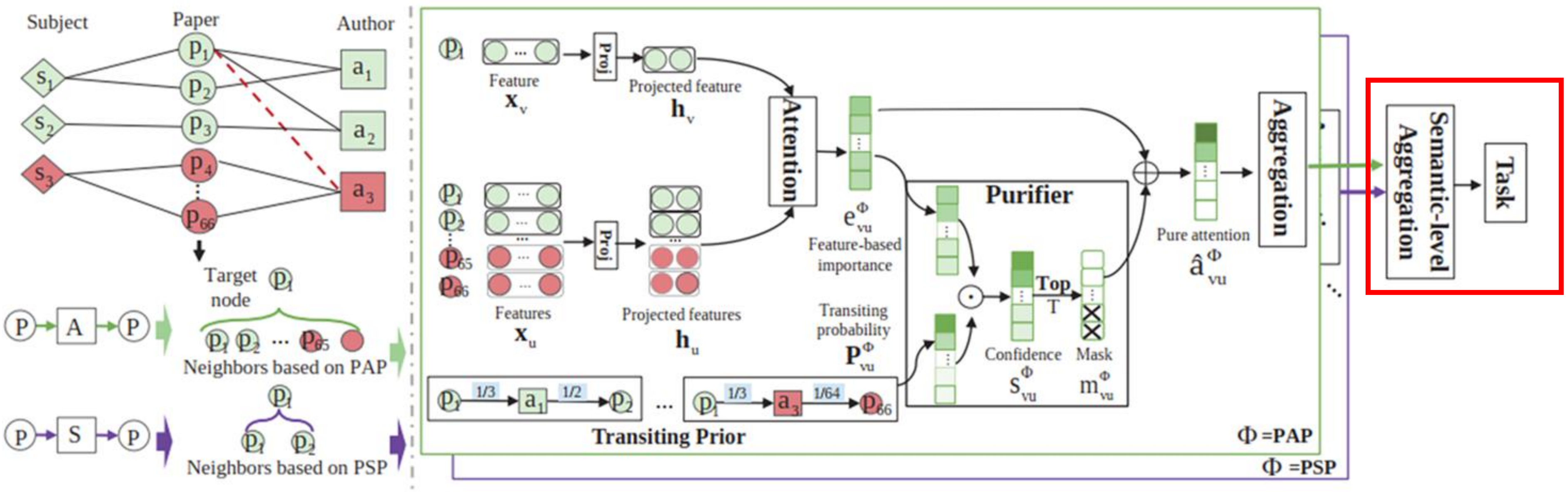
Confidence score. $s_{vu}^{\Phi} = \sigma(\mathbf{P}_{vu}^{\Phi} \cdot e_{vu}^{\Phi}).$ (6)

Purification mask. $m_{vu}^{\Phi} = \begin{cases} \cancel{0}\ 1 & \text{if } u \in \mathrm{Top}(\mathbf{s}_v^{\Phi}, T), \\ -\infty & otherwise, \end{cases}$ (7)

$$\hat{a}_{vu}^{\Phi} = \frac{\exp(m_{vu}^{\Phi} + e_{vu}^{\Phi})}{\sum_{i \in \mathcal{N}_v^{\Phi}} \exp(m_{vi}^{\Phi} + e_{vi}^{\Phi})}. \quad (8)$$

$$\mathbf{z}_v^{\Phi} = \sum_{u \in \mathcal{N}_v^{\Phi}} (\hat{a}_{vu}^{\Phi} \cdot \mathbf{h}_u). \quad (9)$$

Semantic-level Aggregation



$$w^{\Phi} = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \mathbf{q}^T \cdot tanh(\mathbf{W} \cdot \mathbf{z}_v^{\Phi} + \mathbf{b}), \qquad (10)$$

$\mathbf{q}$ is the semantic-level attention vector.

$$\mathbf{z}_v = \sum_{\Phi \in \{\Phi_1, \cdots, \Phi_P\}} \beta^{\Phi} \cdot \mathbf{z}_v^{\Phi}. \qquad (11)$$

uses the softmax function to normalize the importance w^{Φ} to yield the attention value β^{Φ}

$$\mathcal{L} = - \sum_{v \in \mathcal{V}_L} \ln(\mathbf{W}_{clf} \cdot \mathbf{z}_{v, c_v}), \qquad (12)$$

c_v is the class of training node $v \in \mathcal{V}_L$

Experiments

Data	Model	Clean	Attack		
			$\Delta=1$	$\Delta=3$	$\Delta=5$
ACM	HAN	0.926	0.528	0.330	0.240
	Jaccard	0.918	0.892	0.860	0.848
	SimP	0.898	0.746	0.476	0.358
	GGCL	0.902	0.260	0.084	0.084
	HAN-RoHe$_P$	0.924	0.780	0.868	0.870
	HAN-RoHe$_T$	**0.940**	0.900	0.564	0.304
	HAN-RoHe	0.920	**0.904**	**0.902**	**0.882**
DBLP	HAN	0.942	0.332	0.096	0.060
	Jaccard	0.934	0.816	0.812	0.802
	SimP	0.942	0.790	0.670	0.600
	GGCL	0.914	0.684	0.464	0.344
	HAN-RoHe$_P$	0.862	0.686	0.714	0.702
	HAN-RoHe$_T$	**0.944**	0.760	0.360	0.220
	HAN-RoHe	0.942	**0.936**	**0.864**	**0.808**
Aminer	HAN	**0.882**	0.346	0.134	0.102
	GGCL	0.808	0.276	0.056	0.042
	HAN-RoHe$_P$	0.840	0.772	0.772	0.774
	HAN-RoHe$_T$	0.842	0.788	0.668	0.562
	HAN-RoHe	0.838	**0.840**	**0.812**	**0.802**

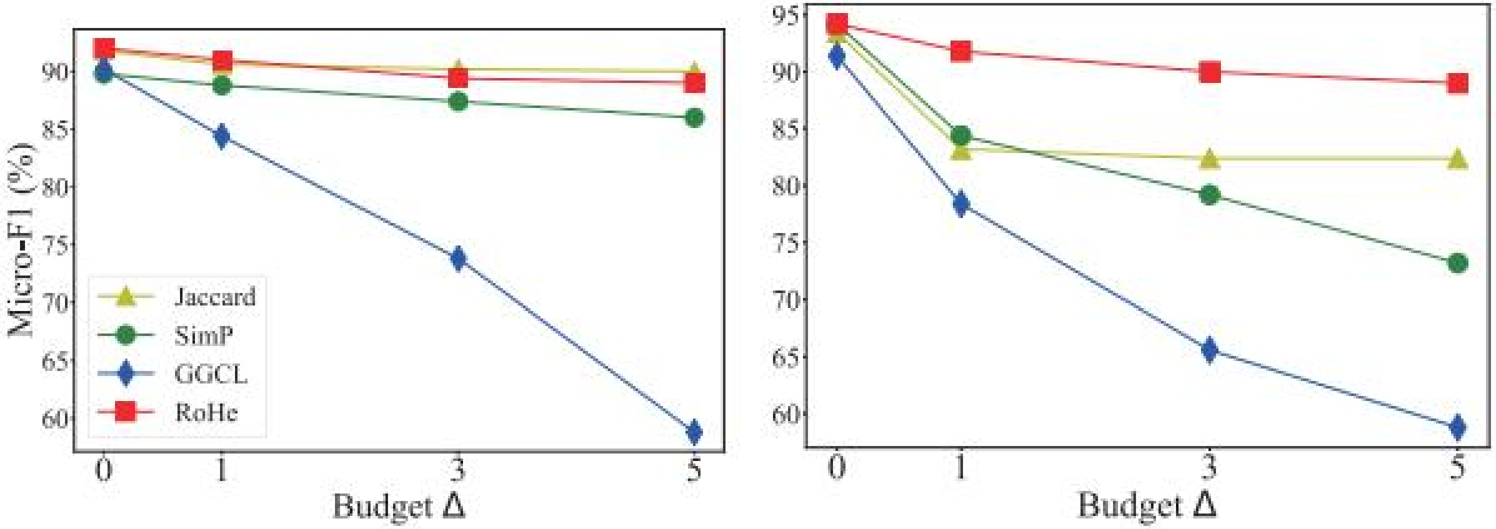
Δ is the maximum number of the perturbed edges

employ FGSM-based attacks to generate perturbation edges

Experiments

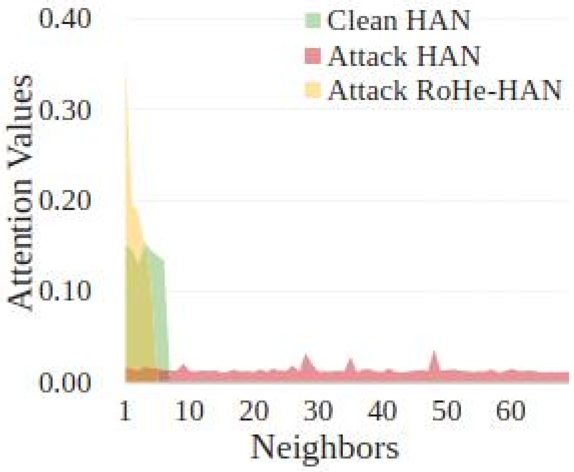
Data	HGNNs	Clean	Attack		
			Δ =1	Δ =3	Δ =5
ACM	HAN	**0.926**	0.528	0.330	0.240
	HAN-RoHe	0.920	**0.904**	**0.902**	**0.882**
	MAGNN	**0.926**	0.711	0.647	0.589
	MAGNN-RoHe	0.916	**0.901**	**0.907**	**0.909**
	GTN	0.932	0.786	0.466	0.302
	GTN-RoHe$_T$	**0.932**	**0.892**	**0.772**	**0.656**
DBLP	HAN	0.942	0.332	0.096	0.060
	HAN-RoHe	**0.942**	**0.936**	**0.864**	**0.808**
	MAGNN	**0.920**	0.620	0.494	0.416
	MAGNN-RoHe	0.898	**0.798**	**0.740**	**0.682**
	GTN	0.946	0.564	0.200	0.128
	GTN-RoHe$_T$	**0.950**	**0.644**	**0.334**	**0.172**

Experiments

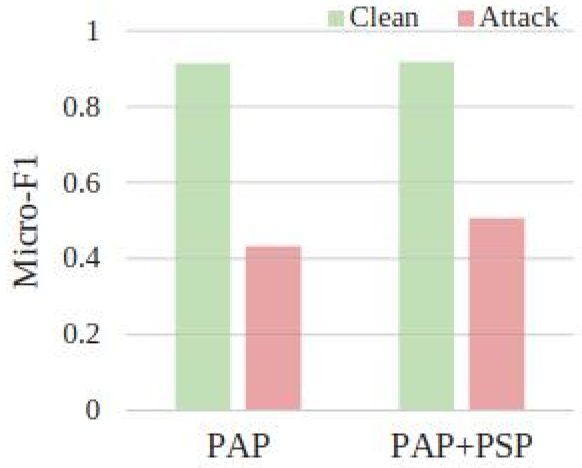


(a) ACM (b) DBLP

Experiments



(a) Node-level



(b) Semantic-level