Additional results

ANONYMOUS

1 F1 SCORE OF PIA ATTACK

Table 1 shows the F1 score of Attack A_1 and A_2. F1 score combines the precision and recall of a classifier into a single metric by taking their harmonic mean. The higher the F1 score means the better performance. It shows the effectiveness of our attack models.

2 DISTRIBUTION OF PIA FEATURES OF POSITIVE AND NEGATIVE GRAPHS

Figure 1 illustrates the distribution of PIA features (the embeddings after max-pooling and posteriors) of graphs with or without property. we can see the significant boundary between these two types of graphs in Figure 1(a), 1(b), 1(c), 1(g), 1(h), which are consistent with the results in Figure2 in our paper that the attack accuracy for these settings are all above 0.97. And another observation is that the distinguishability of link-based property is weak than node-based property, which is consistent with the observation in that the attack accuracy of node-based property mostly are higher than that of link-based property.

3 MORE RESULTS OF INFLUENCE SCORES OF DIFFERENT NODE/LINK GROUP

Table 2 show the influence disparity across different groups on Pokec and Pubmed dataset (we only show the influence disparity on Facebook in the paper), which is to explain why our PIA can work. The values in Table 2 are measures in the same way of Table 6 in the paper, which is described in Section 6.3. From Table 2, we observe that all GNN models have noticeable disparity in the influence scores across different groups. It aligns with our hypothesis that GNNs are "biased" in the sense that they behave differently across different groups in the training data, thus enables GPIA to distinguish positive and negative graphs where different node/link groups dominate these graphs.

Author's address: Anonymous.

settings	GCN						GraphSAGE						GAT					
	P_1	P_2	P_3	P_4	P_5	P_6	P_1	P_2	P_3	P_4	P_5	P_6	P_1	P_2	P_3	P_4	P_5	P_6
A_1^1	0.98	0.96	0.74	0.9	0.53	0.56	0.99	0.86	0.97	0.75	0.86	0.79	0.84	0.76	0.82	0.63	0.64	0.89
A_1^2	0.93	1	0.96	0.85	0.80	0.82	0.98	0.86	0.99	0.74	0.85	0.82	0.74	0.6	0.86	0.62	0.61	0.79
$A_1^{1,2}$	0.98	0.99	0.97	0.9	0.75	0.81	0.99	0.96	0.97	0.75	0.86	0.79	0.84	0.76	0.84	0.63	0.7	0.89
$A_1^{1,2,o}$	0.99	1	0.96	0.85	0.8	0.92	0.98	0.96	0.99	0.74	0.85	0.82	0.84	0.86	0.95	0.62	0.7	0.79
A_2	0.99	1	0.99	0.86	0.78	0.93	0.94	0.96	0.98	0.81	0.76	0.82	0.89	0.91	0.96	0.76	0.72	0.84

Table 1. F1 scores of A_1 and A_2

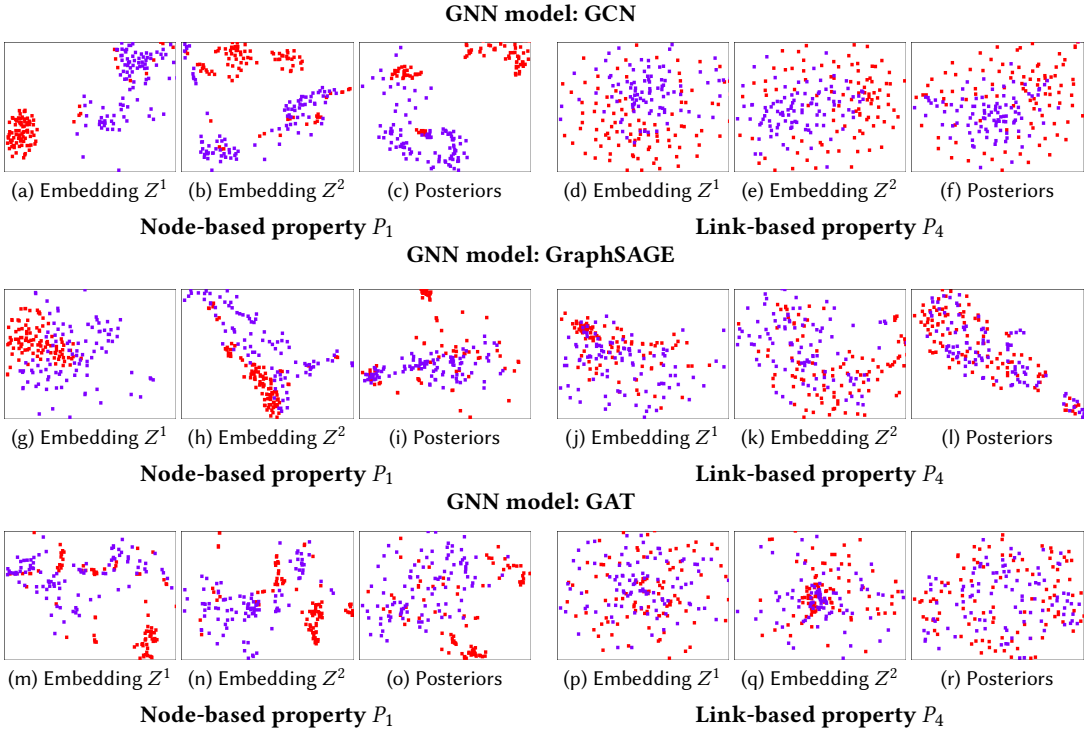

Fig. 1. TSNE visualization of the distribution of PIA input features on Pokec dataset. Blue and red dots denote positive (w/ property) and negative (w/o property) graphs respectively.

GNN model	Pokec dataset				Pubmed dataset			
	Node groups		Link groups		Node groups		Link groups	
	Male	Female	Same-gender	Diff-gender	w/ "IS"	w/o "IS"	"IS"-"IS"	"ST"-""ST""
GCN	0.00853	0.00582	0.02544	0.02147	0.00044	0.00152	0.00128	0.00037
GraphSAGE	0.01447	0.00951	0.03207	0.00666	0.00652	0.01487	0.00337	0.003
GAT	0.00302	0.00104	0.01601	0.00069	0.00171	0.00012	0.00075	0.00124

Table 2. Influence scores of different node/link groups (Pokec and Pubmed dataset).