TED: Related Party Transaction guided Tax Evasion Detection on Heterogeneous Graph

1 Supplementary Material

1.1 Datasets

In the tax heterogeneous graph, the initial feature selection of the company node is shown in Table 1. We select two public datasets from a heterogeneous graph benchmark [1] to verify the effectiveness of our proposed algorithm. The detailed statistics of the four datasets are shown in Table 2, and the degree distribution is shown in Figure 1.

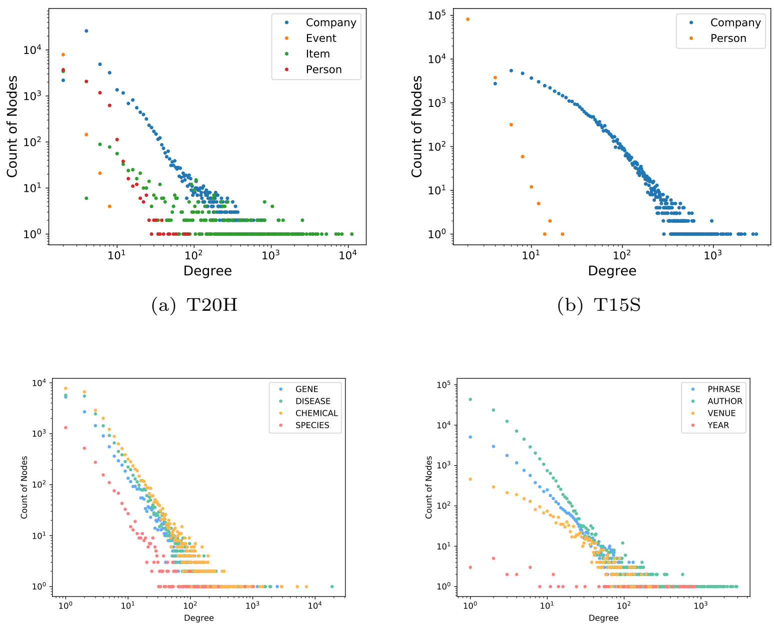
Table 1 Company Feature Selection.

Feature Category	Attribute Symbols	Attribute Type	Attribute Meaning
Registration Information	HYMC	text	industry
	SCCYDM	category	industry code
	DJZCLXMC	text	taxpayer registration type
	NSRZTMC	text	taxpayer current status
	FDDBRNL	number	age of legal representative
	CWFZRNL	number	age of financial officer
	BSRNL	number	age of tax preparers
	JYFW	text	business scope
	CYRS	number	number of people engaged
	ZCZB	number	registration capital
	TZZE	number	total investment
Business Information	XFKPZB	number	sales side invoicing percentage
	GFKPZB	number	purchaser invoicing percentage
	XGFPFSB	number	the ratio of the number of invoices from the seller to the buyer
	FPSYSZB	number	the proportion of upstream companies
	FPXYSZB	number	the proportion of downstream companies
	XSYSB	number	ratio of downstream companies to upstream companies
	XFJEZB	number	seller's amount percentage
	GFJEZB	number	purchaser's amount percentage
	XGFPJEB	number	sales to purchase amount ratio

- **PubMed**. We use PubMed provided by a heterogeneous graph benchmark[1], which contains four node types(i.e., 13561 genes (G), 20163 diseases (D), 26522 chemicals (C), 2863 species (S)) and ten edge types.

Table 2 Statistics of the Datasets.

Dataset	node type	node	edge type	edge	attribute	label type	label node
T20H	4	112015	6	198903	300	2	1770
T15S	2	132522	2	467273	300	2	2072
PubMed	4	63109	10	244986	200	8	454
DBLP	4	206461	6	288959	300	13	618



(a) T20H (b) T15S

(c) PubMed (d) DBLP

Fig. 1 Degree distribution in four real-world tax datasets.

Word2Vec[2] is used to aggregate the word embedding to obtain 200 dimensional features of all node types. The label information divide the diseases into 8 categories, and each labeled disease has only one label.

• **DBLP**. We sample DBLP from the heterogeneous graph benchmark [1], which contains four node types(i.e., 16254 phrases (P), 185048 authors (A), 5076 venues (V), 83 years (Y)) and six edge types. The initial features of papers and phrases are subjected to Word2Vec[2] to aggregate all word embeddings, and the initial features of authors and venues are aggregated according to corresponding paper features. The label information marks a part of the authors from the four research fields into 13 research groups, and each labeled author has only one label.

Table 3 Experimental results (%) of node classification task on PubMed and DBLP datasets.

Datasets	PubMed		DBLP	
	Macro-F1	Micro-F1	Macro-F1	Micro-F1
HIN2Vec	11.41	16.73	15.37	32.86
PTE	9.25	12.11	40.89	50.98
metapath2vec	12.48	14.53	42.91	55.50
TransE	12.84	15.87	36.46	50.49
ConvE	10.31	13.00	40.71	52.44
DistMult	10.02	14.97	18.37	33.50
ComplEx	8.26	14.09	32.87	47.08
HAN	52.03	55.81	42.35	56.41
R-GCN	46.94	48.84	25.44	36.75
HGT	36.38	38.37	32.79	44.44
TED	**57.20**	**59.30**	**47.30**	**58.97**

1.2 Classification

PubMed includes 8 classification tasks and DBLP includes 13 classification tasks. We use the Macro-F1 and Micro-F1 as the evaluation metrics. The results of the node classification task on the PubMed and DBLP data sets are shown in Table 3.

TED still has the best classification effect in PubMed and DBLP. Compared with the sub-optimal, in the PubMed and DBLP datasets, the Macro-F1 increased by 5.17% and 4.39%, and the Micro-F1 increased by 3.49% and 2.56%, respectively.

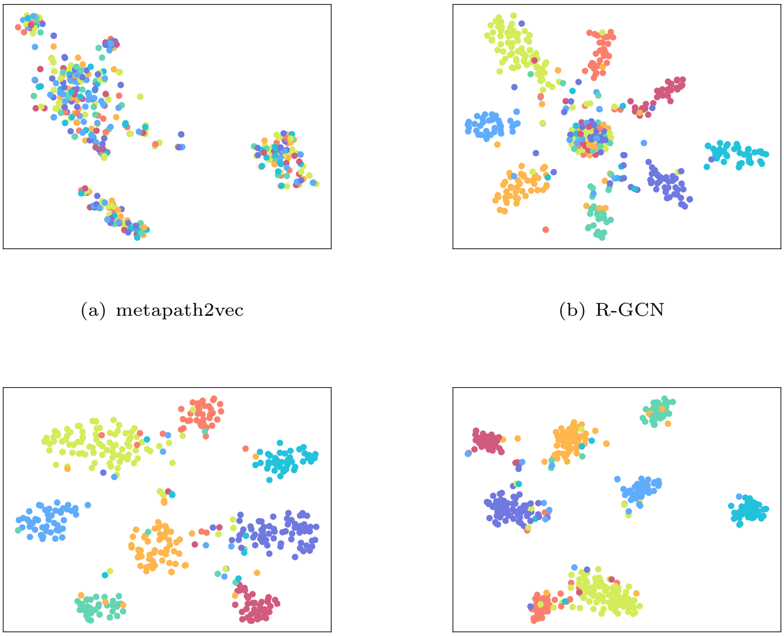
1.3 Clustering

We also conduct clustering experiments to evaluate the performance of different models on clustering tasks. We input the disease node embedding in the PubMed test set and the author node embedding in the DBLP test set into the KMeans algorithm for clustering. Then, we set the number of clusters to the number of label categories, i.e., the number of PubMed clusters is 8, and the number of DBLP clusters is 13. We use the NMI and ARI as the evaluation metrics. Since the performance of KMeans is affected by the initialization, we iteratively repeated the process 10 times and take the average value. The results are shown in Table 4.

As shown in Table 4, TED outperforms all other baselines in clustering tasks. Compared with the sub-optimal, in the PubMed dataset, the NMI increased by 3.65%, and the ARI increased by 4.67%. TED improves by 0.83% on NMI and is suboptimal on ARI.

Table 4 Experimental results (%) of clustering task on PubMed and DBLP datasets.

Datasets	PubMed		DBLP	
	NMI	ARI	NMI	ARI
HIN2Vec	16.96	0.21	29.71	2.11
PTE	13.21	-0.13	34.63	6.99
metapath2vec	16.01	-1.74	41.32	18.58
TransE	14.87	-1.00	42.39	16.79
ConvE	13.34	-1.16	43.30	15.78
DistMult	15.43	-0.78	36.00	9.90
ComplEx	15.24	-1.27	43.23	21.23
HAN	43.15	22.46	54.25	**26.68**
R-GCN	26.97	0.24	28.26	4.09
HGT	27.41	4.48	44.60	15.84
TED	**46.80**	**27.13**	**55.08**	24.51



(a) metapath2vec

(b) R-GCN

(c) HAN

(d) TED

Fig. 2 Visualization on the PubMed dataset. Each node represents a disease, and the color of the node represents the type of disease.

1.4 Visualization

To visualize the node embeddings learned by the model more intuitively, we use t-SNE[3] to project the disease embedding in the PubMed dataset into a two-dimensional space, and we use different colors to represent the corresponding disease categories.

The visualization results are shown in Figure 2, where we show the embedding results of the four models of metapath2vec, R-GCN, HAN and our proposed TED. Figure 2(a) shows the embedding result of metapath2vec. We can see that the embedding is clustered into several clusters, but each cluster contains multiple types of nodes, which does not distinguish different types of nodes. Figure 2(b) is the embedding result of R-GCN. R-GCN basically distinguishes each category, but there is also a cluster containing nodes of various categories, which also explains why R-GCN, a spectral domain method, may have an over-smoothing problem. Figure 2(c) is the embedding result of HAN. The classification result of HAN is very good. There are obvious boundaries between categories, but the distance within the cluster is relatively large. Figure 2(d) is our model, which distinguishes different categories well and has a good degree of similarity within the categories. This also reflects that TED can learn better node embedding.

References

[1] Yang, C., Xiao, Y., Zhang, Y., Sun, Y., Han, J.: Heterogeneous network representation learning: A unified framework with survey and benchmark. TKDE (2020)

[2] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. arXiv preprint arXiv:1310.4546 (2013)

[3] Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)