Adversarial Learning on Heterogeneous Information Networks

Binbin Hu[1, 2], Yuan Fang[3], Chuan Shi[1]

[1]Beijing University of Posts and Telecommunications

[2]AI Department, Ant Financial Services Group

[3]Singapore Management University

KDD2019

Background & Problem

Heterogeneous Information Network (HIN)

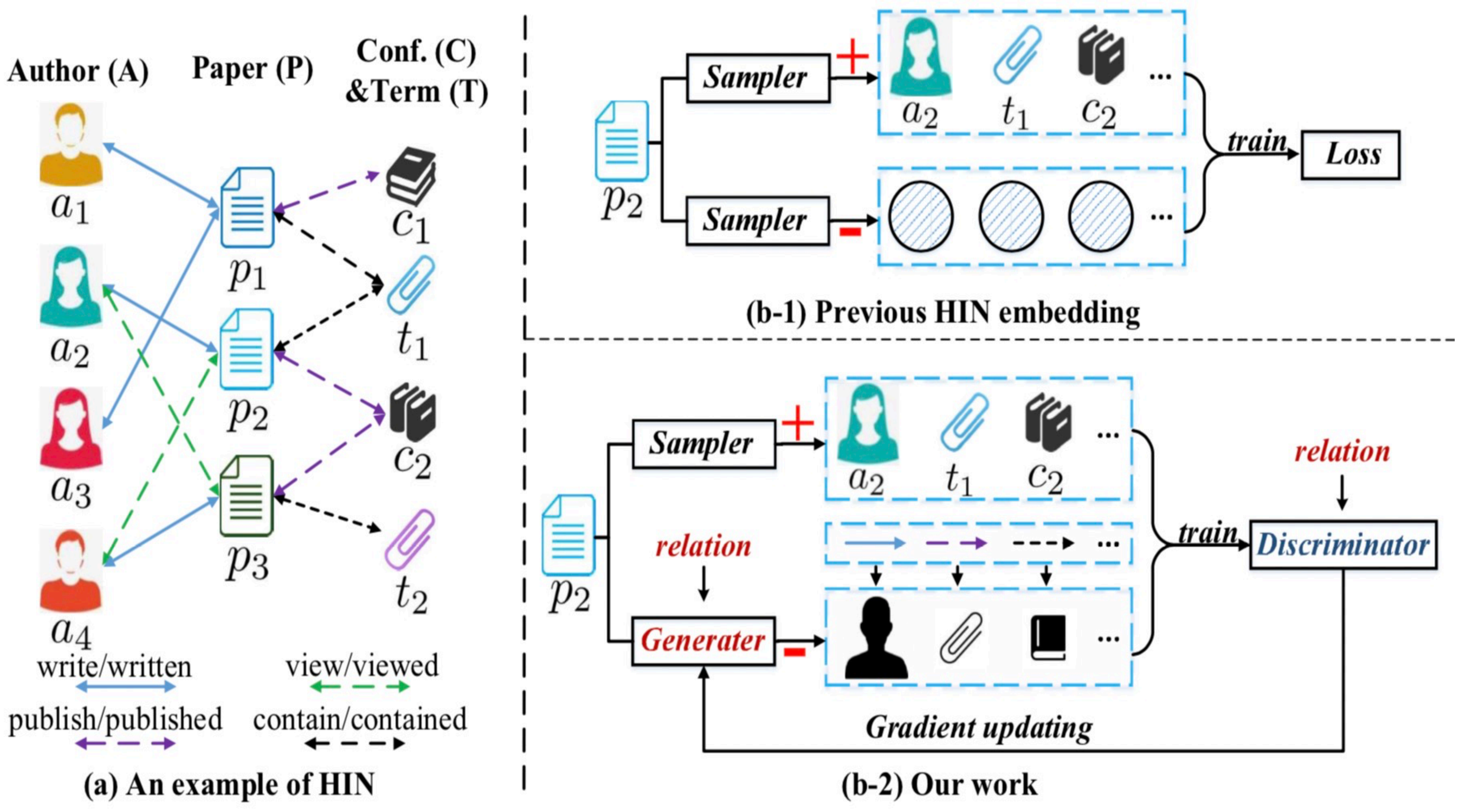
- Include multiple types of nodes and links
- Model heterogeneous data and contain rich semantics

HIN Embedding (HINE)

- Consist of two samplers and one loss function
- Samplers : select positive and negative examples
- Loss function : trained on these samples to optimize node representations

Limitation of HINE

- Randomly select existing nodes in the network as negative samples
- Heed to the latent distribution of the nodes so that lack robustness
- Require domain knowledge



(a) An example of HIN

(b-1) Previous HIN embedding

(b-2) Our work

Adversarial Learning (or GAN)

- Makes the model more robust to sparse or noisy data
- Provides better samples to reduce the labeling requirement
- GraphGAN, ANE, NETRA, ARGA

Limitation on GAN based Embedding

- Only investigate homogeneous networks
- Poor performance on semantic-rich HINs

Our Idea

HINE+ Adversarial Learning

HeGAN : The Proposed Model

HIN Embeddng with GAN based Adversarial Learning (HeGAN)

Challenges

How to capture the semantics of multiple types of nodes and relations?

How to generate fake samples efficiently and effectively?

Solutions

Relation-aware Generator and Discriminator

I. Discriminator can tell whether a node pair is real or fake w.r.t relation

II. Generator can produce fake node pairs that mimic real pairs w.r.t relation

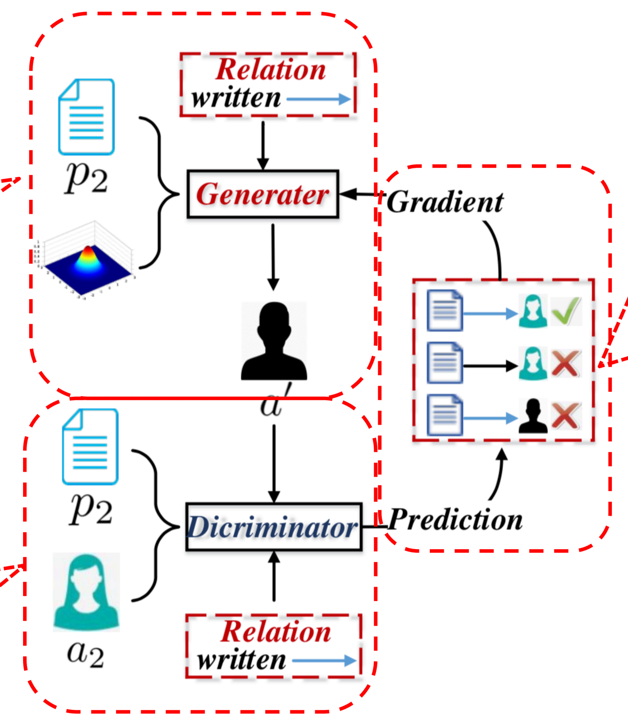
Generalized Generator

I. Sample latent nodes from a continuous distribution

II. No softmax computation and fake samples are not restricted to the existing nodes

Relation-aware, generalized generator

Relation-aware discriminator



Calculate gradient for three cases

I. Connected under given relation

II. Connected under incorrect relation

III. Fake node from relation-aware generator

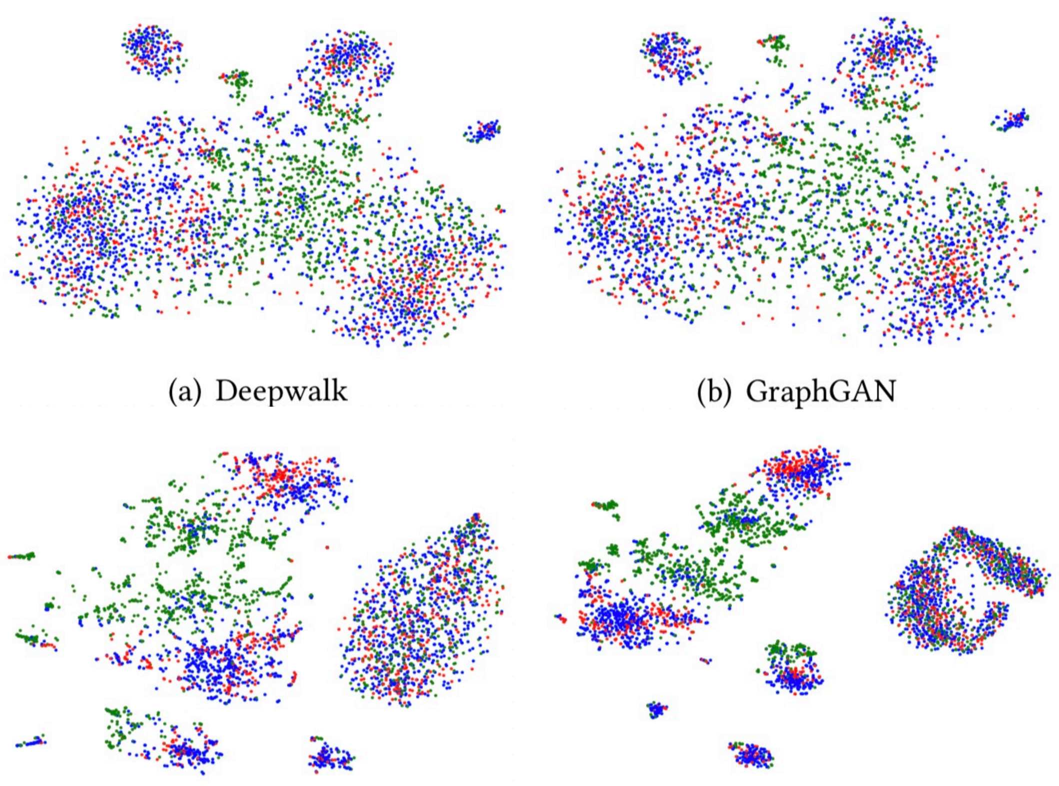
Experiments

Datasets

Datasets	#Nodes	#Edges	#Node types	#Labels
DBLP	37,791	170,794	4	4
Yelp	3,913	38,680	5	3
Aminer	312,776	599,951	4	6
Movielens	10,038	1,014,164	5	N.A.

Node Clustering

Methods	DBLP	Yelp	AMiner
Deepwalk	0.7398	0.3306	0.4773
LINE-1st	0.7412	0.3556	0.3518
LINE-2nd	0.7336	0.3560	0.2144
GraphGAN	0.7409	0.3413	-
ANE	0.7138	0.3145	0.4483
HERec-HNE	0.7274	0.3476	0.4635
HIN2vec	0.7204	0.3185	0.2812
Metapath2vec	0.7675	0.3672	0.4726
HeGAN	**0.7920****	**0.4037****	**0.5052****



(a) Deepwalk (b) GraphGAN (c) Metapath2vec (d) HeGAN

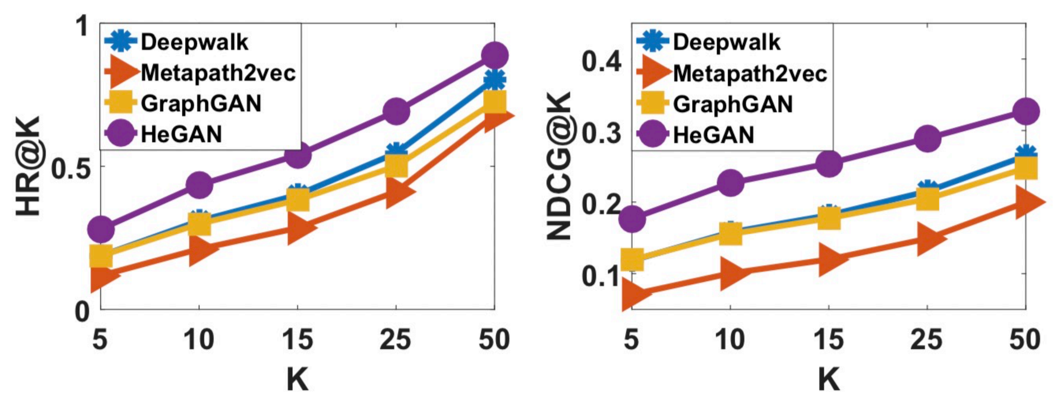
Node Classification

Methods	DBLP			Yelp			AMiner		
	Micro-F1	Macro-F1	Accuracy	Micro-F1	Macro-F1	Accuracy	Micro-F1	Macro-F1	Accuracy
Deepwalk	0.9201	0.9242	0.9298	0.8262	0.7551	0.8145	0.9519	0.9460	0.9529
LINE-1st	0.9239	0.9213	0.9285	0.8229	0.7440	0.8126	0.9776	0.9713	0.9788
LINE-2nd	0.9144	0.9172	0.9236	0.7591	0.5518	0.7571	0.9469	0.9341	0.9471
GraphGAN	0.9198	0.9210	0.9286	0.8098	0.7268	0.7820	-	-	-
ANE	0.9143	0.9153	0.9189	0.8232	0.7623	0.7932	0.9256	0.9203	0.9221
HERec-HNE	0.9214	0.9228	0.9299	0.7962	0.7713	0.7912	0.9801	0.9726	0.9784
HIN2vec	0.9141	0.9115	0.9224	0.8352	0.7610	0.8200	0.9799	0.9775	0.9801
Metapath2vec	0.9288	0.9296	0.9360	0.7953	0.7884	0.7839	0.9853	0.9860	0.9857
HeGAN	**0.9381****	**0.9375****	**0.9421****	**0.8524****	**0.8031****	**0.8432****	**0.9864***	**0.9873***	**0.9883***

Link Prediction

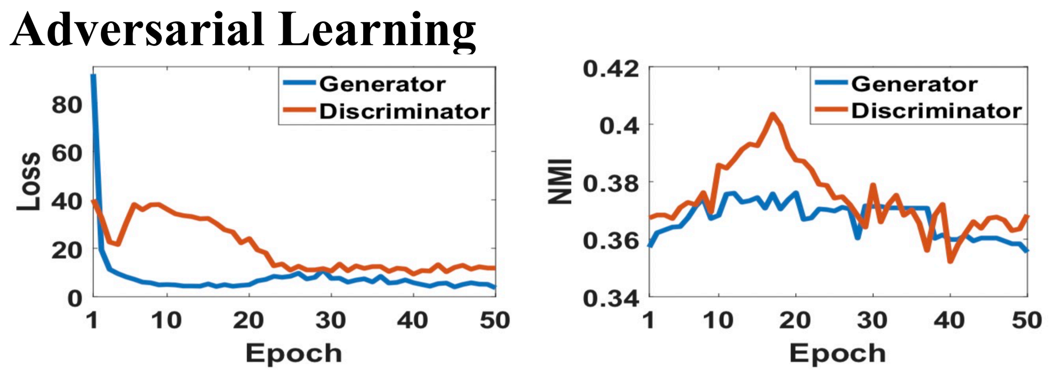
	Methods	DBLP			Yelp			AMiner		
		Accuracy	AUC	F1	Accuracy	AUC	F1	Accuracy	AUC	F1
Logistic Regression	Deepwalk	0.5441	0.5630	0.5208	0.7161	0.7825	0.7182	0.4856	0.5182	0.4618
	LINE-1st	0.6546	0.7121	0.6685	0.7226	0.7971	0.7099	0.5983	0.6413	0.6080
	LINE-2nd	0.6711	0.6500	0.6208	0.6335	0.6745	0.6499	0.5604	0.5114	0.4925
	GraphGAN	0.5241	0.5330	0.5108	0.7123	0.7625	0.7132	-	-	-
	ANE	0.5123	0.5430	0.5280	0.6983	0.7325	0.6838	0.5023	0.5280	0.4938
	HERec-HNE	0.7123	0.7823	0.6934	0.7087	0.7623	0.6923	0.7089	0.7776	0.7156
	HIN2vec	0.7180	0.7948	0.7006	0.7219	0.7959	0.7240	0.7142	0.7874	0.7264
	Metapath2vec	0.5969	0.5920	0.5698	0.7124	0.7798	0.7106	0.7069	0.7623	0.7156
	HeGAN	**0.7290****	**0.8034****	**0.7119****	**0.7240****	**0.8075****	**0.7325****	**0.7198****	**0.7957****	**0.7389****
Inner Product	Deepwalk	0.5474	0.7231	0.6874	0.5654	0.8164	0.6953	0.5309	0.6064	0.6799
	LINE-1st	0.6647	0.7753	0.7363	0.6769	0.7832	0.7199	0.6113	0.6899	0.7123
	LINE-2nd	0.4728	0.4797	0.6325	0.4193	0.7347	0.5909	0.5000	0.4785	0.6666
	GraphGAN	0.5532	0.6825	0.6214	0.5702	0.7725	0.6894	-	-	-
	ANE	0.5218	0.6543	0.6023	0.5432	0.7425	0.6324	0.5421	0.6123	0.6623
	HERec-HNE	0.5123	0.7473	0.6878	0.5323	0.6756	0.7066	0.6063	0.6912	0.6798
	HIN2vec	0.5775	0.8295	0.6714	0.6273	**0.8340**	0.4194	0.5348	0.6934	0.6824
	Metapath2vec	0.4775	0.6926	0.6287	0.5124	0.6324	0.6702	0.6243	0.7123	0.6953
	HeGAN	**0.7649****	**0.8712****	**0.7837****	**0.7391****	0.8298	**0.7705****	**0.6505****	**0.7431****	**0.7752****

Recommendation



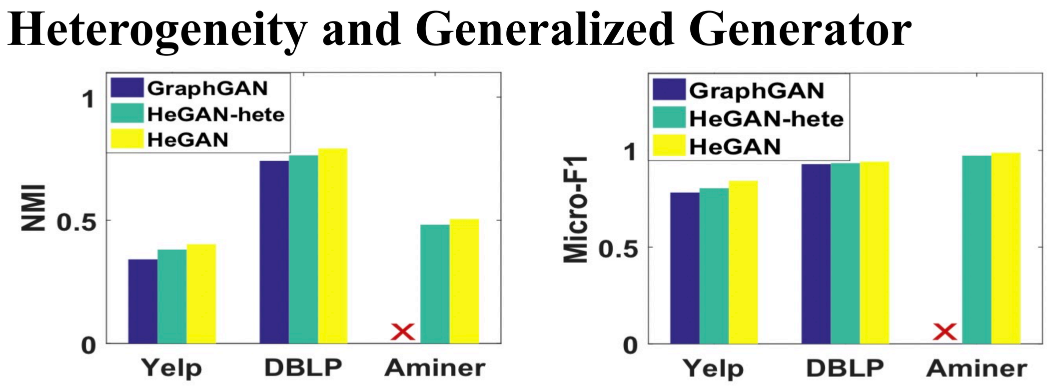
(a) HR (b) NDCG

Adversarial Learning



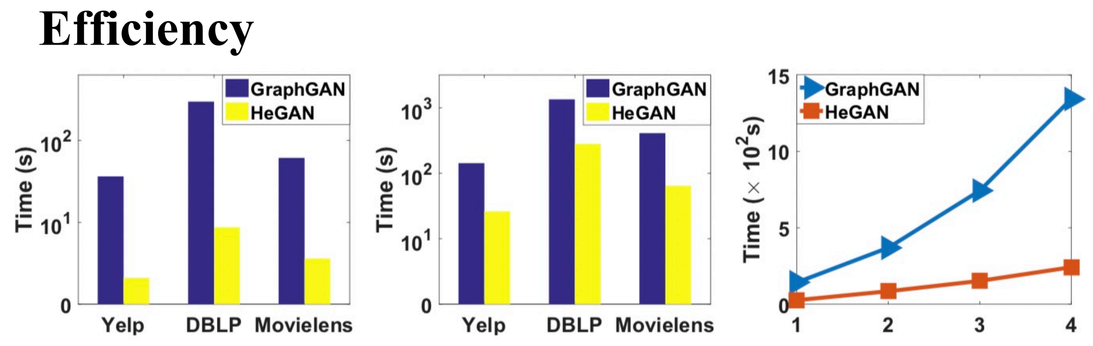
(a) Loss change (b) Performance change in node clustering

Heterogeneity and Generalized Generator



(a) Node clustering (b) Node classification

Efficiency



(a) Generator training (b) Overall training (c) Scalability

Conclusions

- We are the first to employ adversarial learning for HIN embedding, in order to utilize the rich semantics on HINs
- We propose HeGAN that is not only relation-aware to capture rich semantics, but also equipped with a generalized generator
- Extensive experimental results have revealed the effectiveness and efficiency of HeGAN

Acknowledgements

This research was supported by the National Natural Science Foun- dation of China (No. 61772082, 61702296), the National Key Research and Development Program of China (2017YFB0803304), the Beijing Municipal Natural Science Foundation (4182043) and the Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 grant (Approval No. 18-C220-SMU-006).

Contact

- hubinbin@bupt.edu.cn
- (bin.hbb@antfin.com)
- shichuan@bupt.edu.cn
- yfang@smu.edu.sg