Pre-training on Large-Scale Heterogeneous Graph

Xunqiang Jiang[1], Tianrui Jia[1], Yuan Fang[2],
Chuan Shi[1,3*], Zhe Lin[3], Hui Wang[3]

[1]Beijing University of Posts and Telecommunications

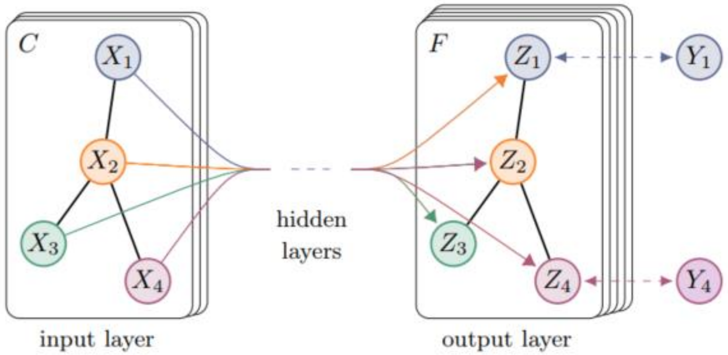
[2]Singapore Management University

[3]Peng Cheng Laboratory, Shenzhen, China

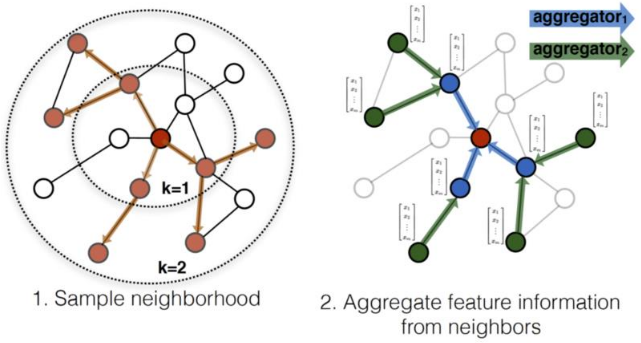
CONTENTS

CONTENTS

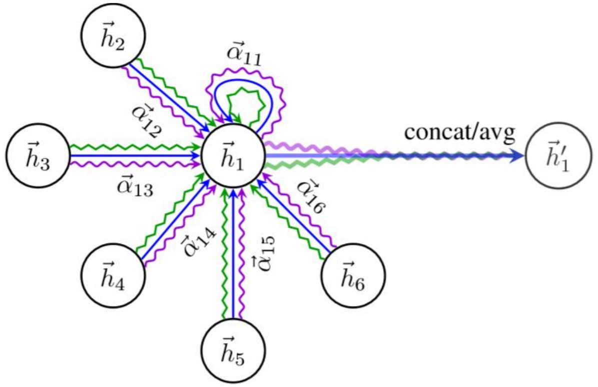
Graph Neural Network (GNN)



GCN

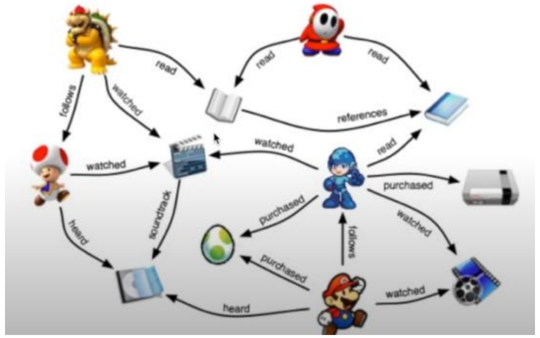


GraphSAGE

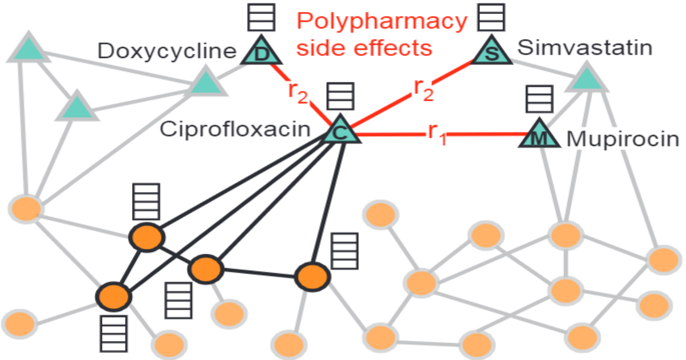


GAT

Various Applications of GNN



Recommendation System



Polypharmacy

• • • • • •

Label Scarcity on Graph Data

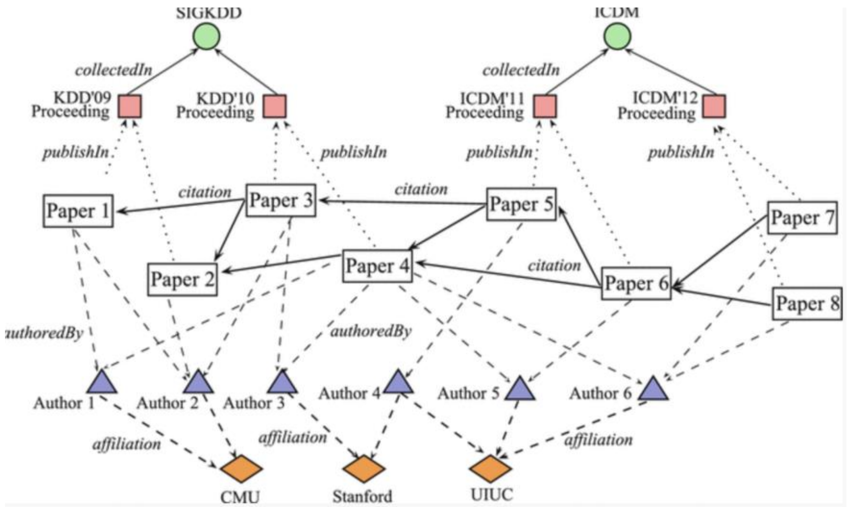
- **GNNs need abundant task-specific labeled → Better results**
  - However, labeled data is usually **expensive** or **infeasible** to obtain

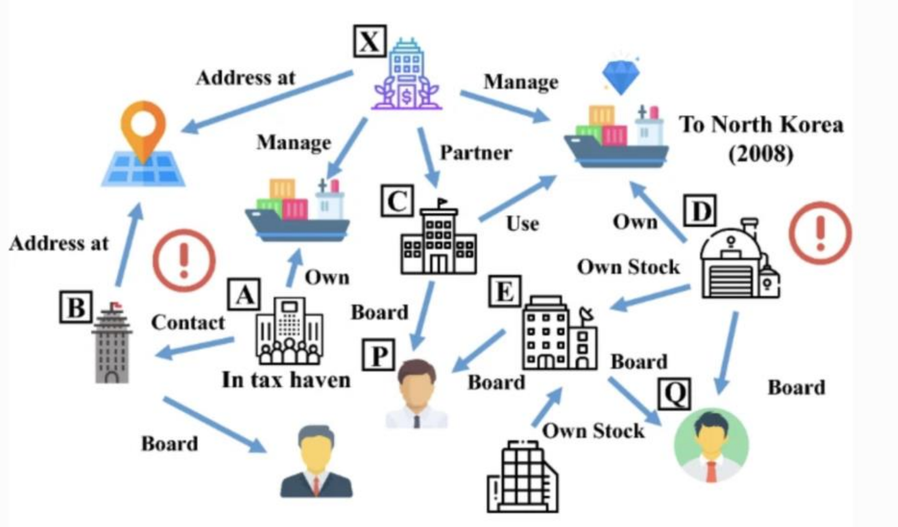
- **Learning from Unlabeled Data → Pre-training**
  - **Unlabeled** data (i.e., the whole graph) is abundant
  - Recent progresses of pre-training in CV and NLP relieve the **reliance** on labeled data, and some recent works propose to pre-train GNNs in a **self-supervised** manner

Existing pre-training methods for GNNs

- They are mainly designed for **homogeneous** graphs

Heterogeneous Graphs





Large heterogeneous graphs with **different relations** and **rich semantics**

Two fundamental problems

- **1. How to effectively capture the semantic and structural properties on a heterogeneous graph during pre-training**
  - Structural properties, rich semantics …… → **varying characteristics** of different types
  - Preserve the inherent semantic and structural properties → **Node** and **Network Schema**

- **2. How to efficiently pre-train GNNs on a large-scale heterogeneous graph**
  - Real-word heterogeneous graphs : billions of nodes and edges → **Scalability**

CONTENTS

PT-HGNN

Preserve **heterogeneous** semantic and structural properties as transferable knowledge, and sparsify **large-scale** heterogeneous graph for efficient pre-training

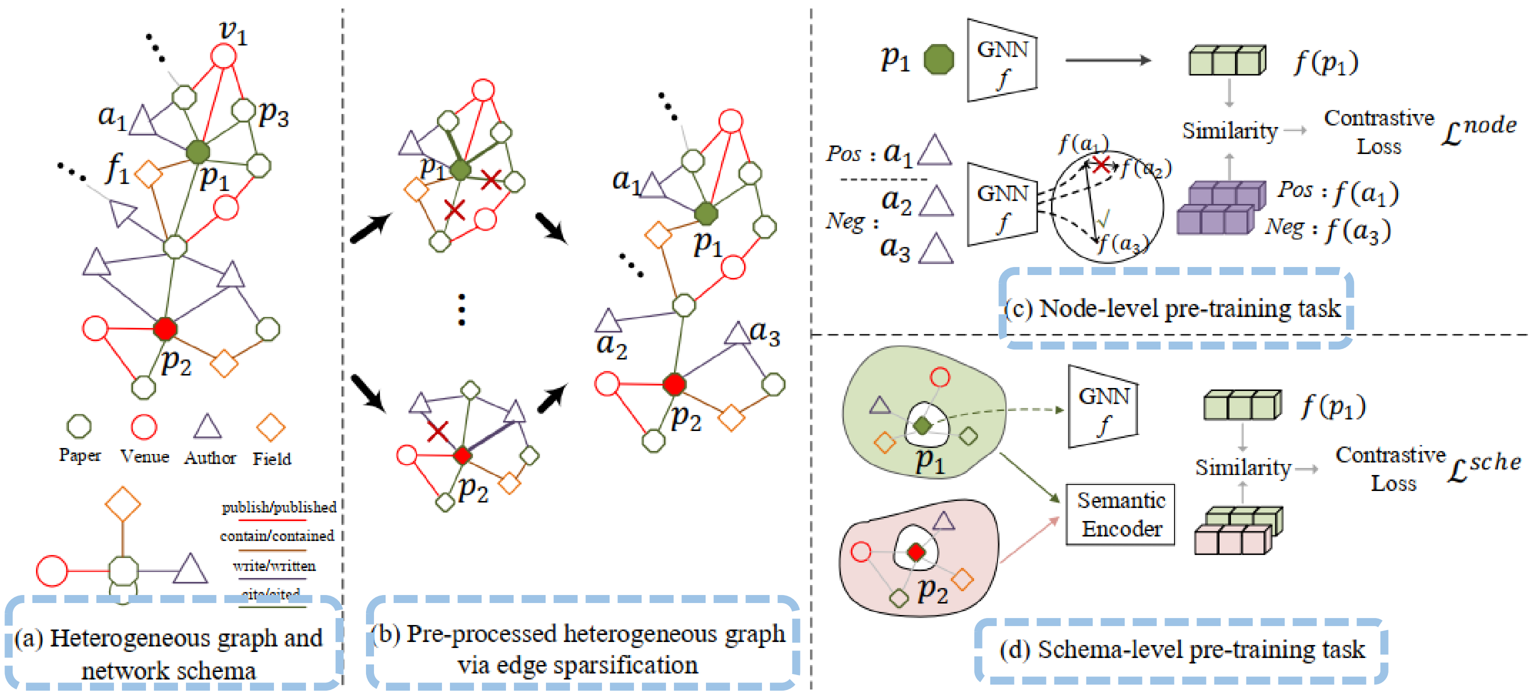
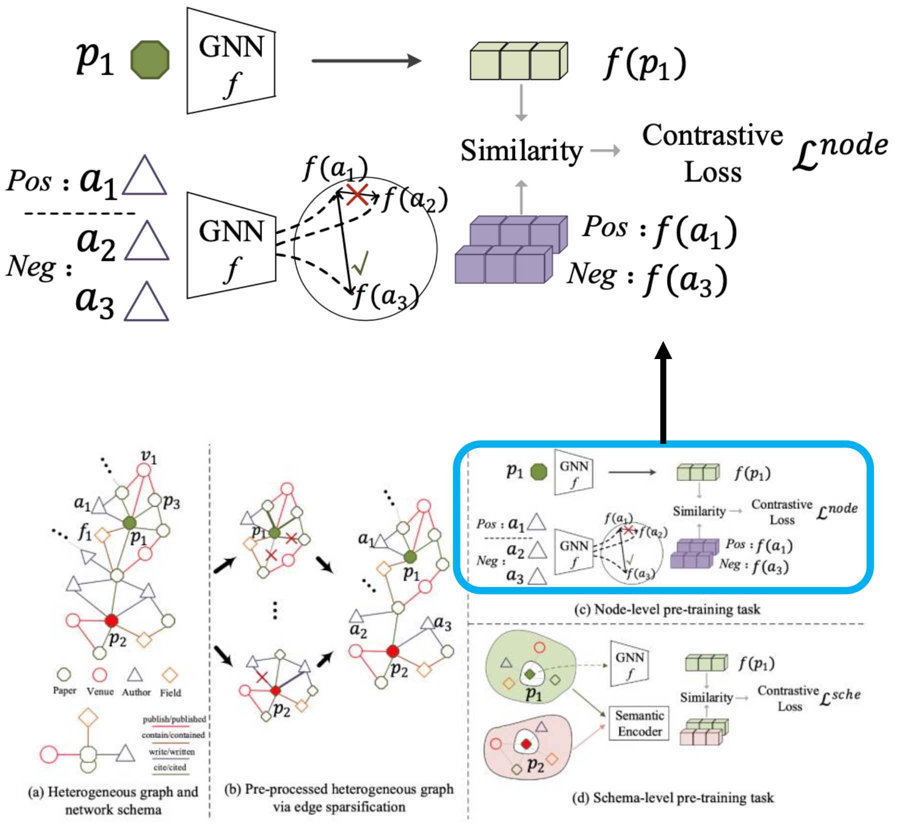


Figure 1: The overall framework of PT-HGNN.

- Relation-based sparsification for efficiency
- Design the node- and schema-level pre-training tasks

Node-level pre-training task: Negative sample selection



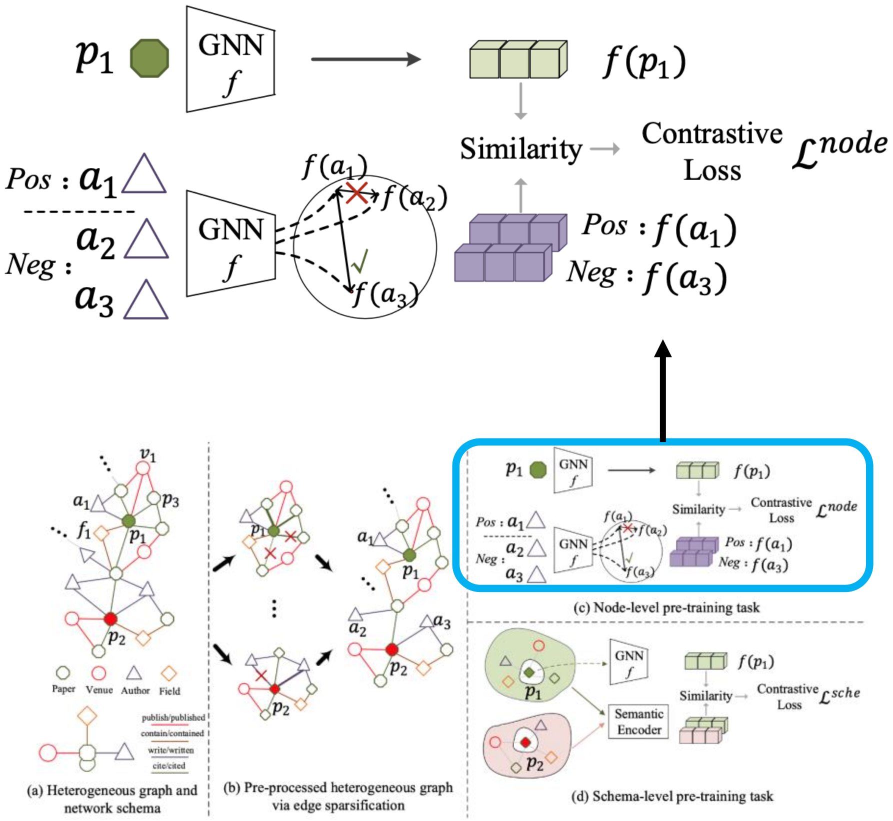
$$\mathcal{N}^{node}_{\langle u,R,v\rangle} = \{\langle u,R,v^-\rangle \mid \overset{1}{\boxed{\phi(v)=\phi(v^-),}}(u,v^-)\notin\mathcal{E},\overset{2}{\boxed{Sim(v,v^-)\le\delta\}}}$$

1. Select negative samples from the **same relation**

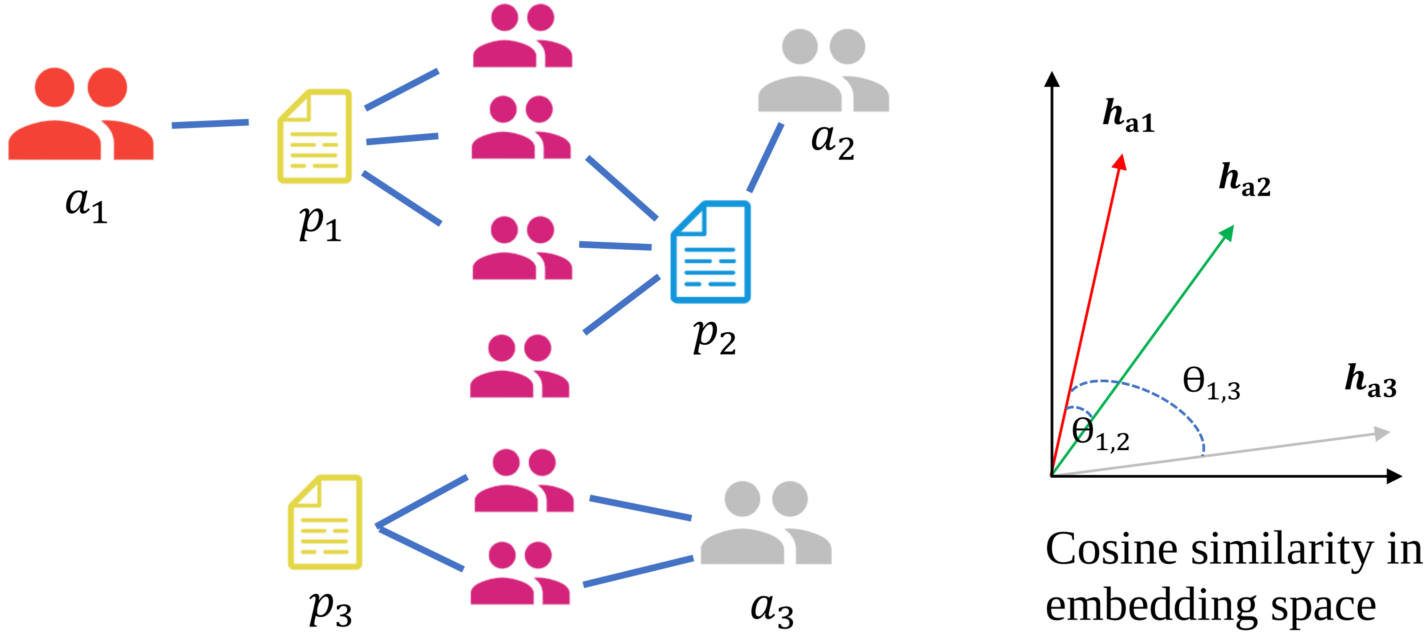
<author1, “writes”, paper1> → <author1, “writes”, paper2> ✓

<author1, “writes”, paper1> → <author1, “writes”, institute1> ✗

Node-level pre-training task: Negative sample selection



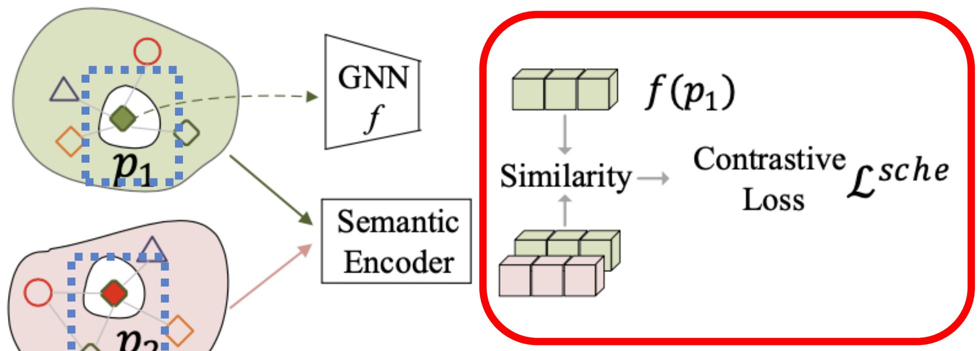
2. Select negative samples **dissimilar enough**

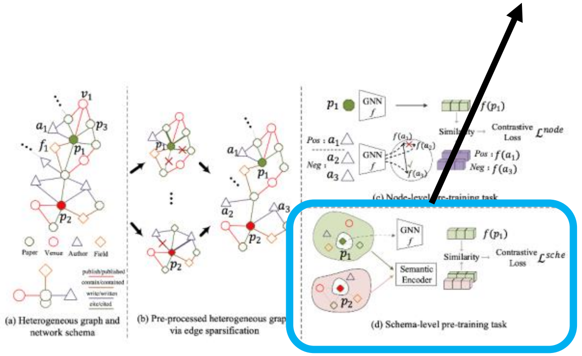


Cosine similarity in the embedding space

For paper $\boldsymbol{p_1}$ and relation paper-author, $\boldsymbol{a_1}$ is the positive sample, $\boldsymbol{a_2}$ is a similar sample, $\boldsymbol{a_3}$ is a negative sample.

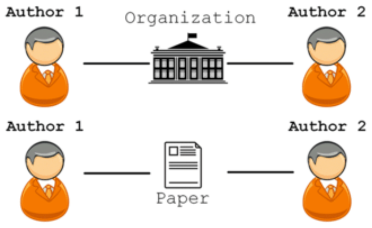
Schema-level pre-training task



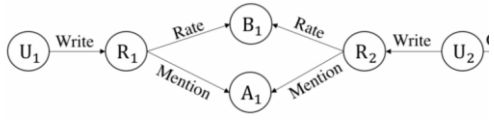


Model the relation between **center node** with **context nodes**

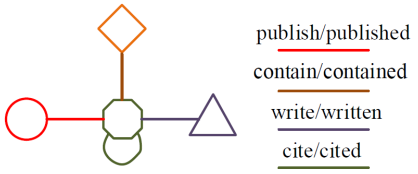
- **Network schema** captures both high-order semantics and structural properties



Meta-path



Motif

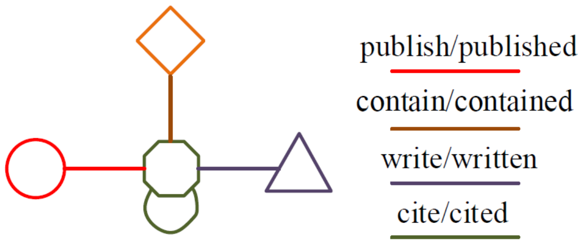


schema

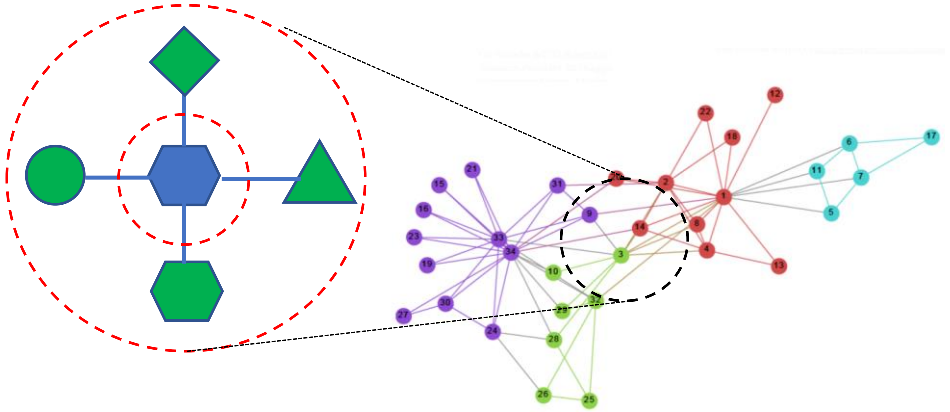
- meta-graph is limited to express high-order structure
- motif is intractable to match when the graph is so large
- schema is the only defining structure that captures both semantic and structural properties

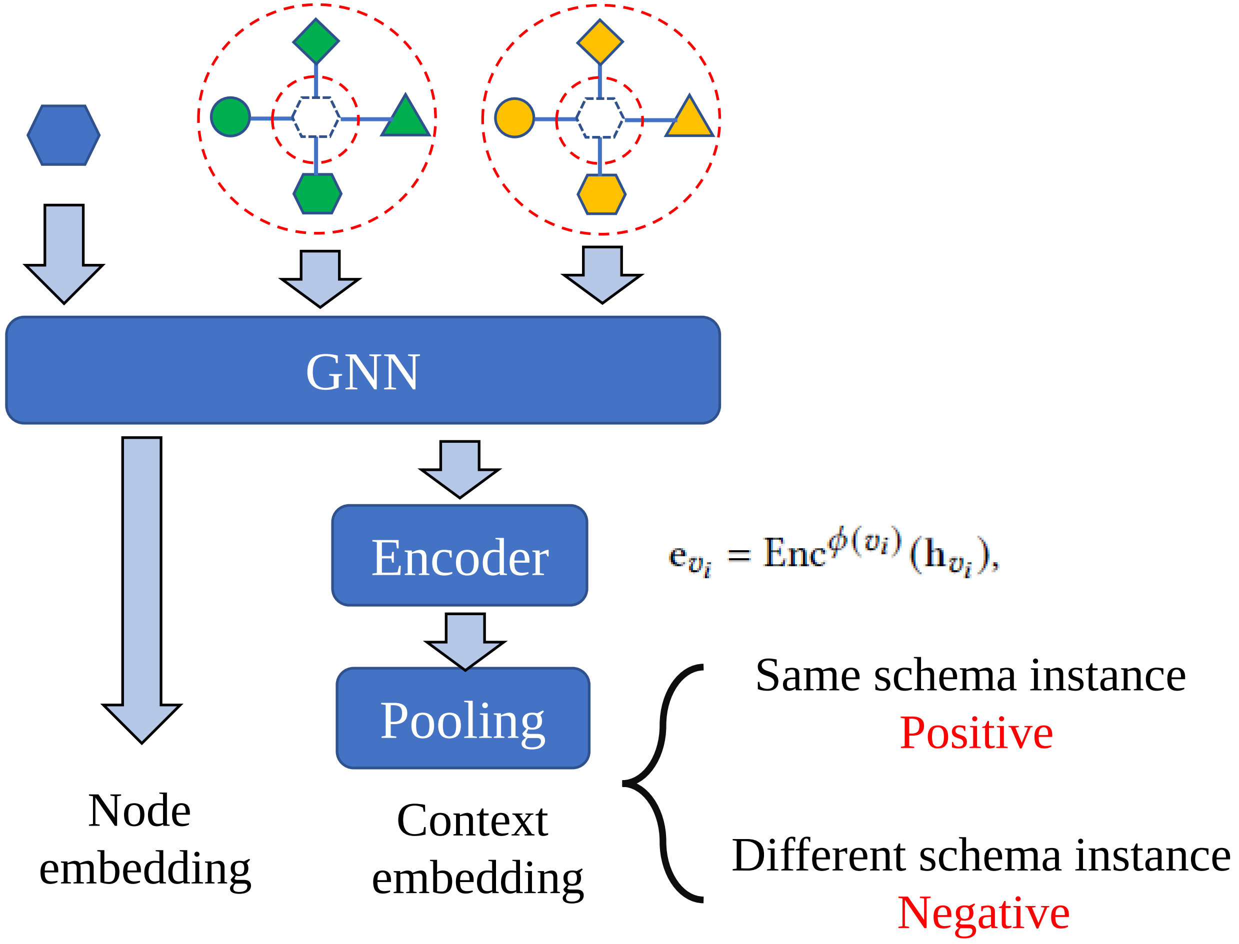
Schema-level pre-training task

sampling according to schema



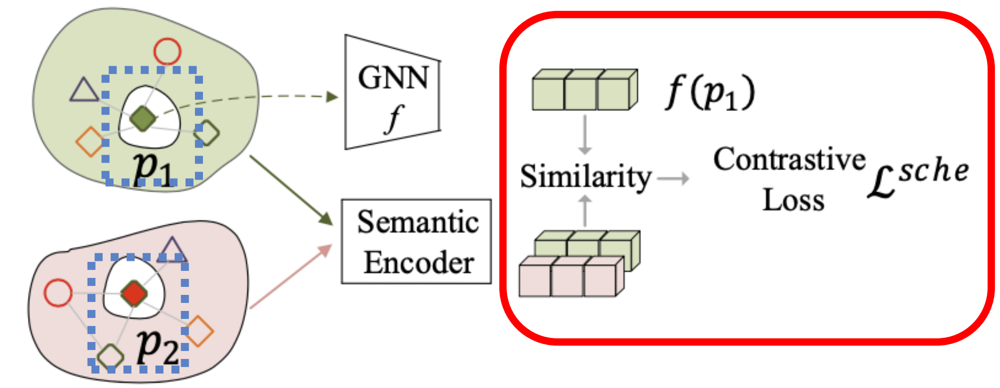
center node with **context nodes**

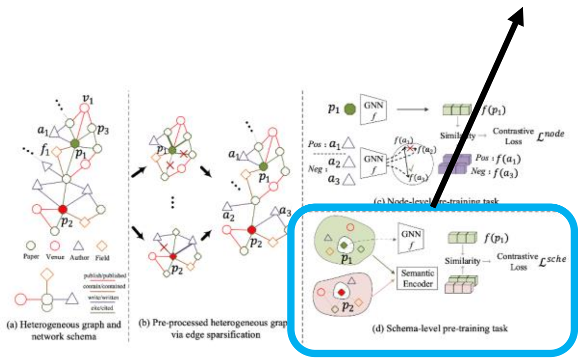




$$\mathrm{e}_{v_i} = \mathrm{Enc}^{\phi(v_i)}(\mathrm{h}_{v_i}),$$

Schema-level pre-training task





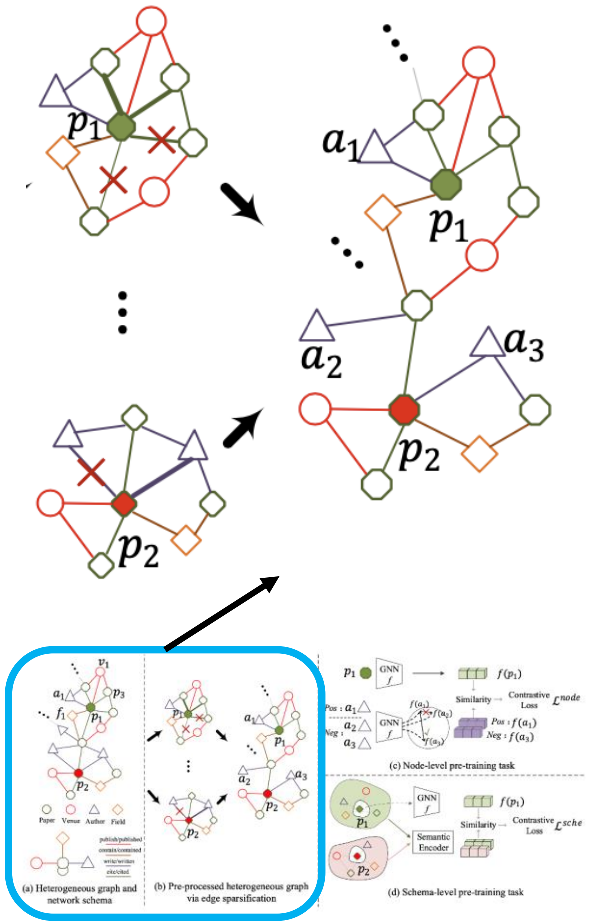
- Sampling Negative samples from:
  - **The current batch** with two target nodes of the **same type**

$$\mathcal{N}_u^1 = \{\mathcal{P}_{u^-}^{sche} \mid u^- \in \mathcal{V}_B, u \neq u^-, \boxed{\phi(u) = \phi(u^-)}\}$$

  - **Previous batch** with two target nodes of the **same type**

$$\mathcal{N}_u^2 = \{\mathcal{P}_v^{sche} \mid \phi(u) = \phi(v), v \in \mathcal{V}_B^{t-1}\},$$

Edge Sparsification



Why edge Sparsification:

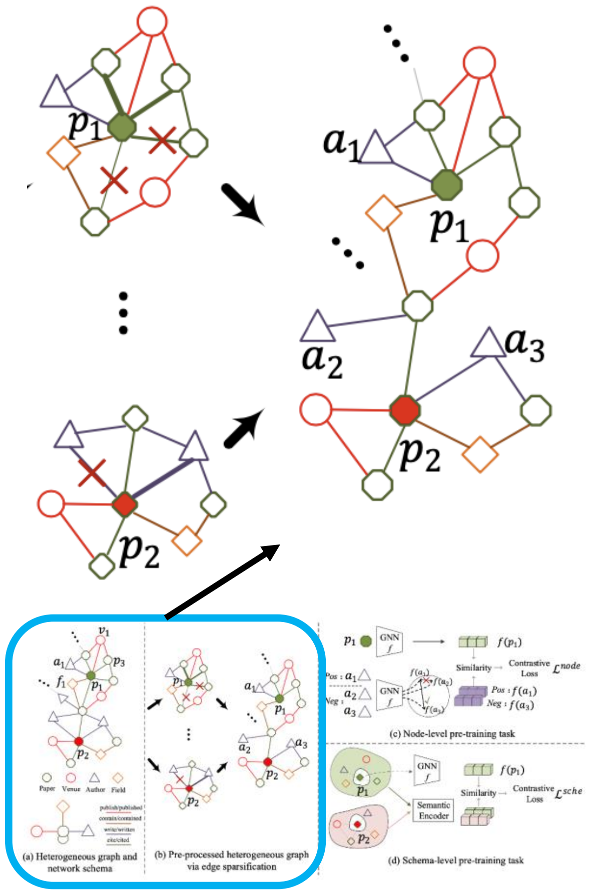
- Preserve more **meaningful** edges (lower noise in graphs)
- Improve the **time efficiency** on large graph

Method : **Relation based Personalized PageRank**

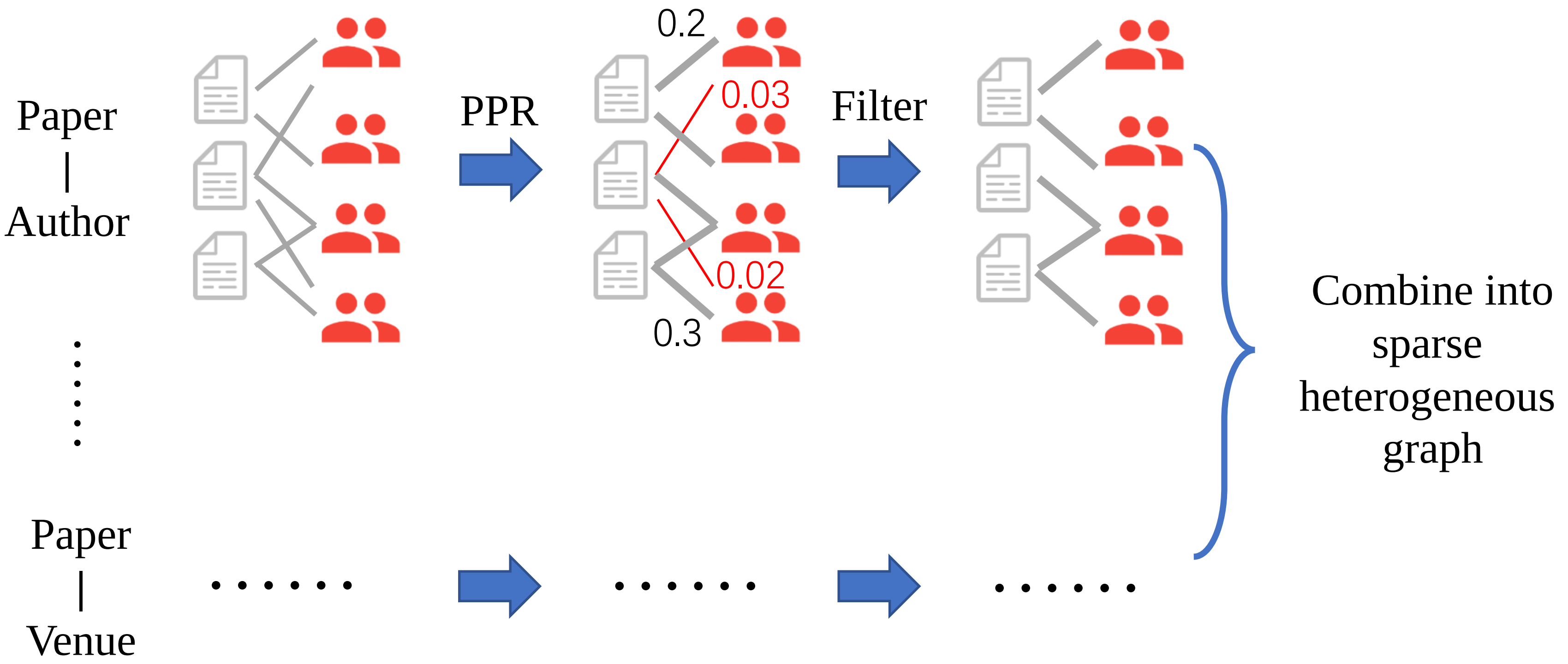
Acceleration :

Random-Walk Formulation (Forward Search) + Top-K Entries

Edge Sparsification



The construction process of sparse heterogeneous graph



CONTENTS

Baselines

- EdgePred
- DGI
- ContextPred
- GPT-GNN
- GraphCL
- No-Pretrain

Datasets

Open Academic Graph (OAG) unifies two academic graphs:
Microsoft Academic Graph and Aminer

Statistics of Datasets

Dataset	#nodes	#edges	#venues	#papers	#fields	#authors	#institutes	#P-V	#P-P	#P-F	#P-A	#A-I
CS	11,918,983	107,263,811	27,433	5,597,605	289,930	5,985,759	18,256	5,597,606	31,441,552	47,462,559	15,571,614	7,190,480
Mater	4,552,941	42,161,581	15,141	2,442,235	79,305	2,005,362	10,898	2,442,235	13,011,272	19,119,947	5,582,765	2,005,362
Engin	5,191,920	36,146,719	19,867	3,239,504	99,444	1,819,100	14,005	3,239,504	4,848,158	22,498,822	3,741,135	1,819,100
Chem	12,158,967	159,537,437	19,142	7,193,321	183,782	4,748,812	13,910	7,193,321	74,018,600	57,162,528	16,414,176	4,748,812
OAG	178,663,987	2,236,196,802	53,073	89,606,257	615,288	88,364,081	25,288	89,606,258	1,021,237,518	657,049,405	300,853,688	167,449,933

#nodes: 178 million; #edges: 2 billion

Experiment results on Node classification and Link Prediction

Whole graph: OAG
Domain specific subgraphs: Computer science, material, chemistry, engineering

Dataset	Downstream Task		No pre-train	EdgePred	DGI	ContextPred	GraphCL	GPT-GNN	PT-HGNN	Improv.
CS	Paper–Field	NDCG	27.42±0.42	31.37±0.32	32.82±0.67	33.15±0.71	32.64±0.65	<u>35.24±0.47</u>	**36.04**±0.37	2.27%
		MRR	23.17±0.45	32.13±0.52	33.43±0.81	33.24±0.57	33.24±0.67	<u>33.57±0.71</u>	**37.76**±0.42	12.48%
	Paper–Venue	NDCG	27.76±0.56	35.77±0.59	34.23±0.71	34.30±0.92	32.11±0.69	<u>36.15±0.53</u>	**38.81**±0.51	7.35%
		MRR	11.39±0.37	16.34±0.47	16.21±0.62	17.66±0.81	16.29±0.49	<u>19.13±0.65</u>	**21.19**±0.45	10.76%
	Author ND	NDCG	76.27±0.53	79.41±0.68	81.38±0.93	79.22±0.72	79.95±0.89	<u>80.20±0.51</u>	**82.19**±0.60	2.48%
		MRR	54.82±0.49	59.06±0.74	58.98±0.79	60.23±0.83	60.55±0.74	<u>60.94±0.52</u>	**63.38**±0.38	4.00%
OAG	Paper–Venue	NDCG	42.28±0.50	43.25±0.61	<u>44.23±0.53</u>	43.07±0.74	42.66±0.66	44.05±0.75	**47.13**±0.68	6.56%
		MRR	22.76±0.37	23.40±0.35	24.38±0.35	24.12±0.42	25.03±0.48	<u>25.19±0.45</u>	**26.75**±0.57	6.19%
	Author ND	NDCG	76.52±1.13	78.01±0.86	77.98±0.93	77.88±0.72	78.11±0.93	<u>79.33±0.87</u>	**79.99**±0.92	0.83%
		MRR	54.65±0.53	58.00±0.63	58.30±0.48	57.49±0.60	58.22±0.53	<u>59.08±0.52</u>	**61.32**±0.55	3.79%

On average 4.98% improvement: our proposed pre-training strategy is capable of exploiting transferable information and graph properties on heterogeneous graphs

Experiment results on Node classification and Link Prediction

Evaluate the effect of node- and schema-level pre-training tasks on heterogeneous graphs

Downstream Task		No pre-train	PT-HGNN$_{node}$	PT-HGNN$_{sche}$	PT-HGNN
Paper–Field	NDCG	27.42	35.80	35.16	36.04
	MRR	23.17	36.82	36.21	37.76
Paper–Venue	NDCG	27.76	36.23	35.24	38.81
	MRR	11.39	20.42	18.92	21.19
Author ND	NDCG	76.27	80.41	81.25	82.19
	MRR	54.82	60.57	62.02	63.38

- Node-level : PT−HGNN$_{node}$
- Schema-level : PT−HGNN$_{sche}$
- Combination : PT−HGNN

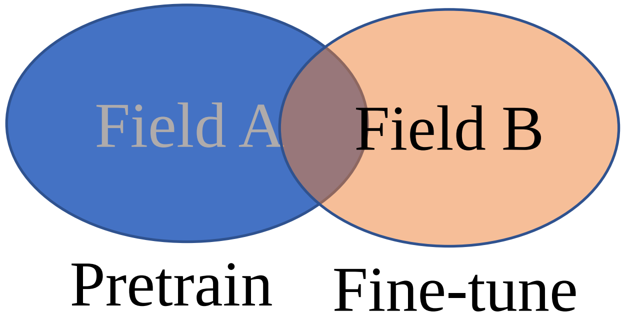
- In link prediction, PT−HGNN$_{node}$ model the pairwise interaction, which performs better
- In node classification, PT−HGNN$_{sche}$ obtain better performance by focusing on modeling the structure context
- The combination offers strong capability in both downstream tasks

Freezing vs. Full Fine-tuning

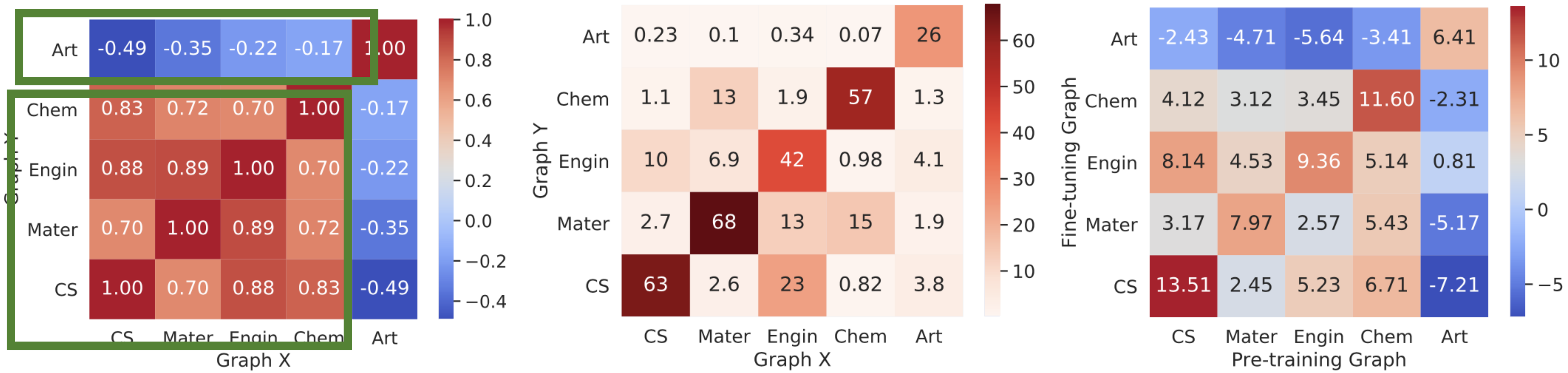
Downstream Task		No pre-train	PT-HGNN (FE)	PT-HGNN
Paper–Field	NDCG	27.42	32.81	36.04
	MRR	23.17	32.50	37.76
Paper–Venue	NDCG	27.76	35.93	38.81
	MRR	11.39	18.45	21.19
Author ND	NDCG	76.27	81.41	82.19
	MRR	54.82	62.15	63.38

- PT-HGNN(FE) achieves better performance than the no pre-train model, which are able to capture the **transferable knowledge**.
- the performance of PT-HGNN in freezing mode exhibits **competitive performance** to that of the full fine-tuning mode in some cases

Transfer Experiment



- Knowledge transferring from pre-training to fine-tuning **does not guarantee** a gain in performance
- Positive correlation value between graphs results in **positive transferring** and vice versa



(a) Correlation metric computed with a series of graph properties.

(b) The citation coefficient: the percentage of publications in graph Y that have citations in graph X

(c) MRR gain (%) of the proposed method over the method with no pre-training

Time Efficiency

Downstream Task		No PPR	PPR	Improv.
Paper–Field	NDCG	**36.54**	36.04	-1.38%
	MRR	**38.12**	37.76	-0.95%
Paper–Venue	NDCG	37.82	**38.81**	2.62%
	MRR	20.42	**21.19**	3.77%
Author ND	NDCG	80.87	**82.19**	1.63%
	MRR	60.09	**63.38**	5.48%
Time Efficiency	Time Per Batch(s)	64.2	**37.9**	41.97%

- With the edges sparsification based on personalized PageRank, the training efficiency is increased
- Pre-training on the pre-processed heterogeneous graph achieve the competitive performance

CONTENTS

Conclusion

- **PT-HGNN**, which is a pre-training framework, enables the GNN to capture heterogeneous semantics and structural properties
- Edge sparsification strategy retains meaningful graph structures while accelerating the pre-training procedure
- Extensive experiments on one of the largest heterogeneous graphs

Virtual Conference

KDD2021

August 14th – 18th

Thank you!

Xunqiang Jiang, BUPT, skd621@bupt.edu.cn

More Information:

http://www.shichuan.org/

http://www.yfang.site