Contrastive Pre-Training of GNNs on Heterogeneous Graphs

Xunqiang Jiang[1], Yuanfu Lu[2], Yuan Fang[3], Chuan Shi[1*]

[1]Beijing University of Posts and Telecommunications

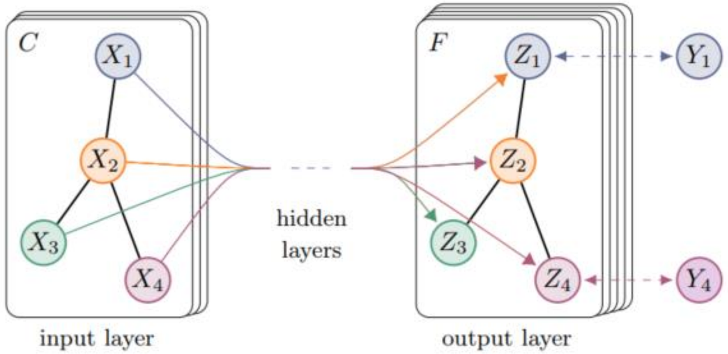
[2]WeChat Search Application Department

[3]Singapore Management University

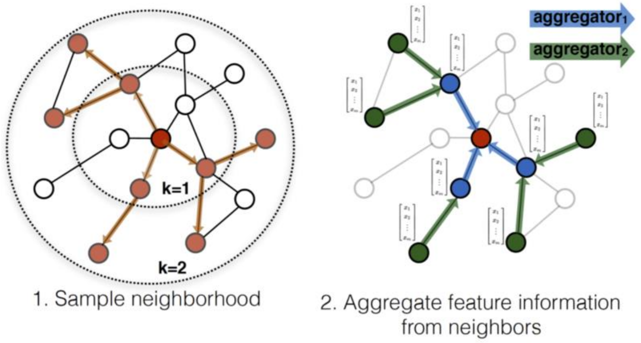
CONTENTS

CONTENTS

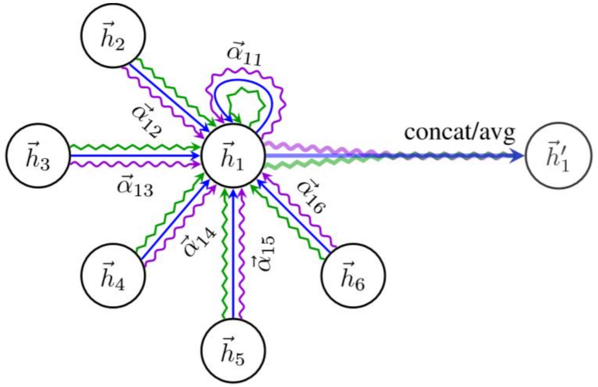
Graph Neural Network (GNN)



GCN

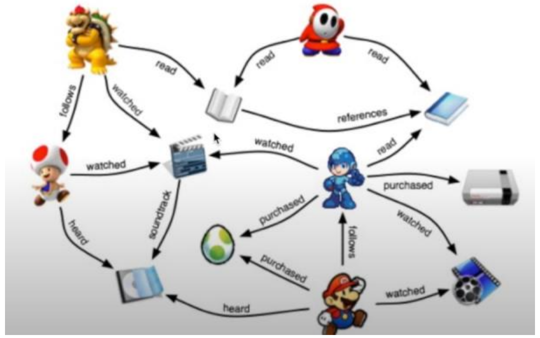


GraphSAGE

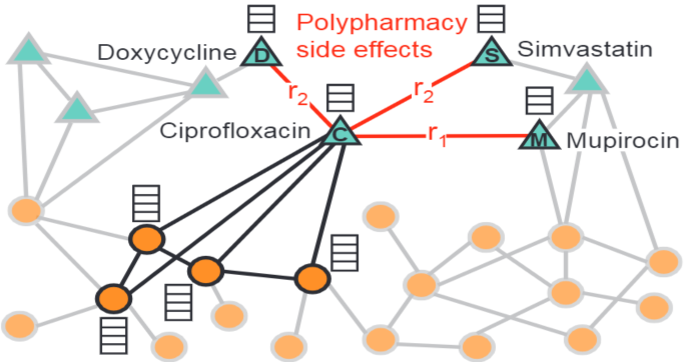


GAT

Various Applications of GNN



Recommendation System



Polypharmacy

……

Label Scarcity on Graph Data

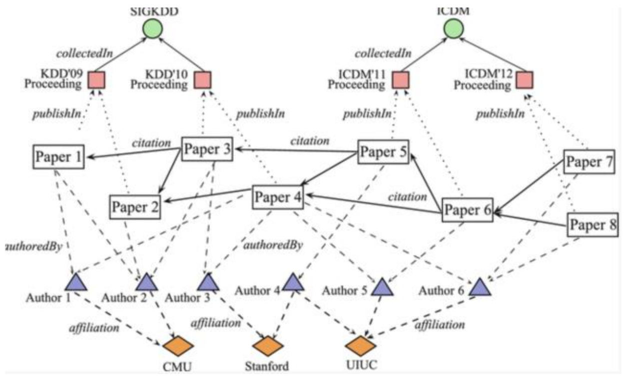
- **GNNs need abundant task-specific labeled → Better results**
  - However, labeled data is usually **expensive** or **infeasible** to obtain

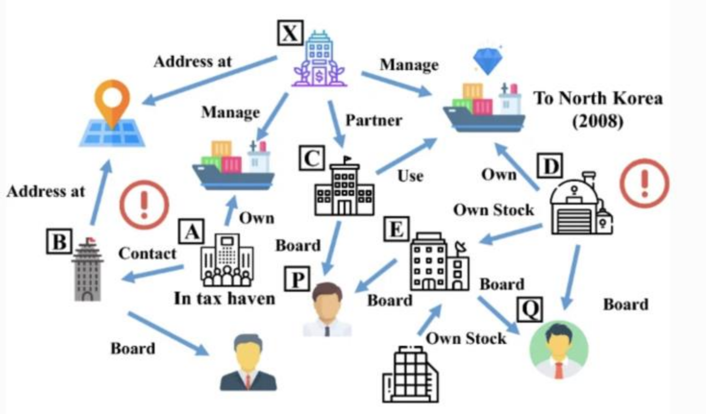
- **Learning from Unlabeled Data → Pre-training**
  - **Unlabeled** data (i.e., the whole graph) is abundant
  - Recent progresses of pre-training in CV and NLP relieve the **reliance** on labeled data, and some recent works propose to pre-train GNNs in a **self-supervised** manner

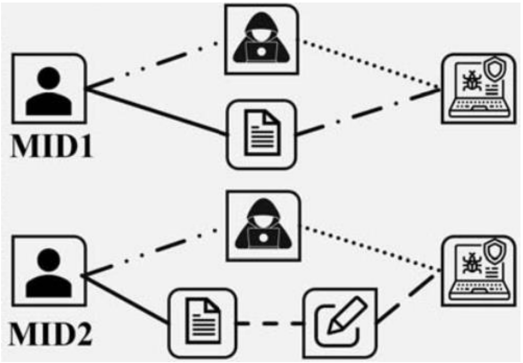
Existing pre-training methods for GNNs

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

Heterogeneous Graphs and meta graph







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

Two fundamental problems

- **1. How to distinguish and tailor to different types of nodes and edges during pre-training**
  - Defining characteristics of a heterogeneous graph …… → different node and edge types

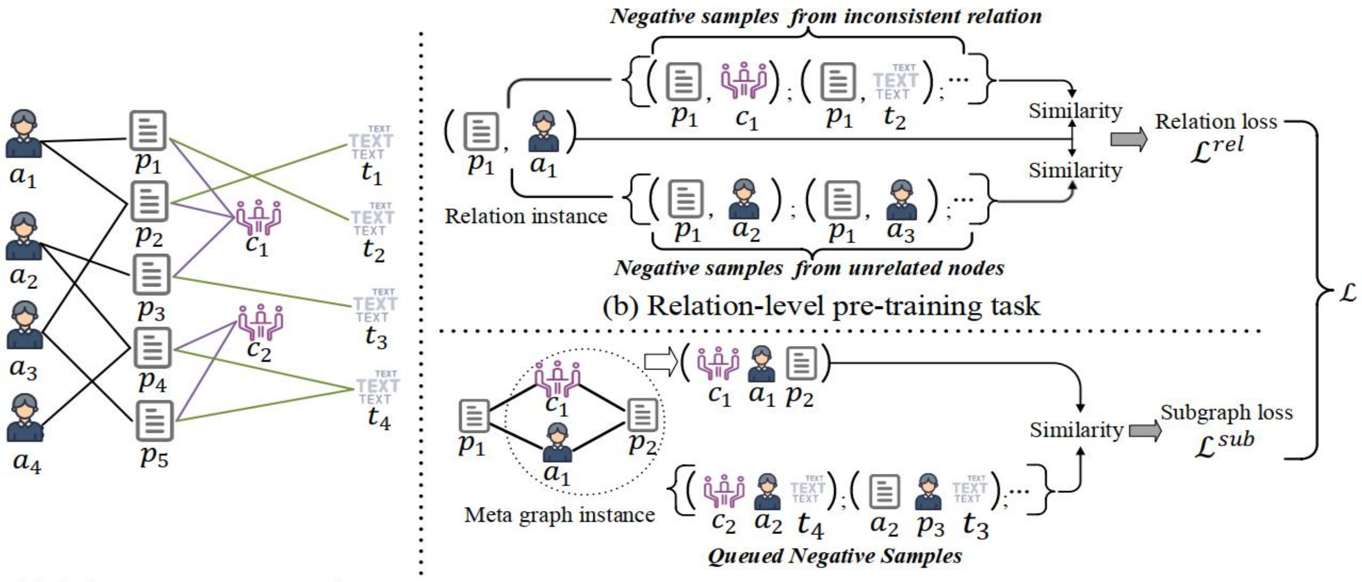
- **2. How to further preserve high-order semantic contexts during pre-training**
  - High-order semantics → advanced structures for pre-training (i.e., meta-graph)

CONTENTS

Overall architecture

CPT-HG

Leverage contrastive pre-training tasks on relation- and subgraph-level, which captures both the semantic and structural properties for pre-training GNNs on heterogeneous graph



(a) A heterogeneous graph

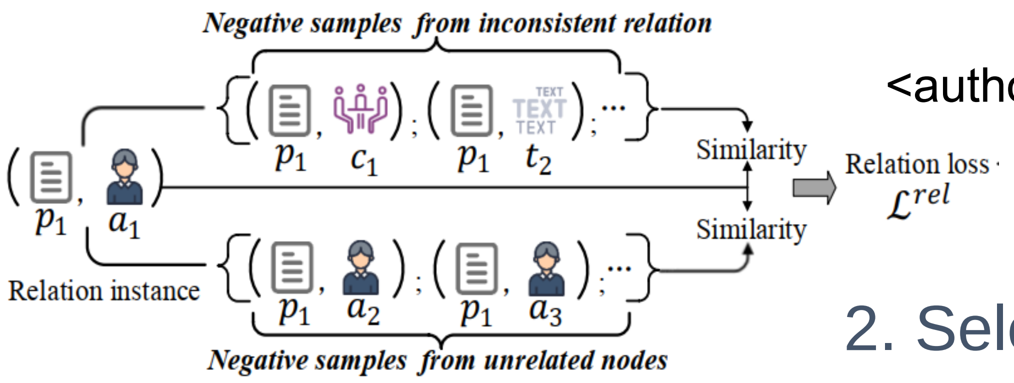
(b) Relation-level pre-training task

(c) Subgraph-level pre-training task

- Relation-level: differentiate the relation type between two nodes
- Subgraph-level: captures high-order structure and embodies complex multiparty relationships

Relation-level Pre-training Task

Encode the relational semantics which constitute the basis of heterogeneity



(b) Relation-level pre-training task

1. Select negative samples from the **unrelated node**

<author1, “writes”, paper1>

<author1, “writes”, paper2>

<author1, “writes”, institute1>

2. Select negative samples from the **inconsistent relation**

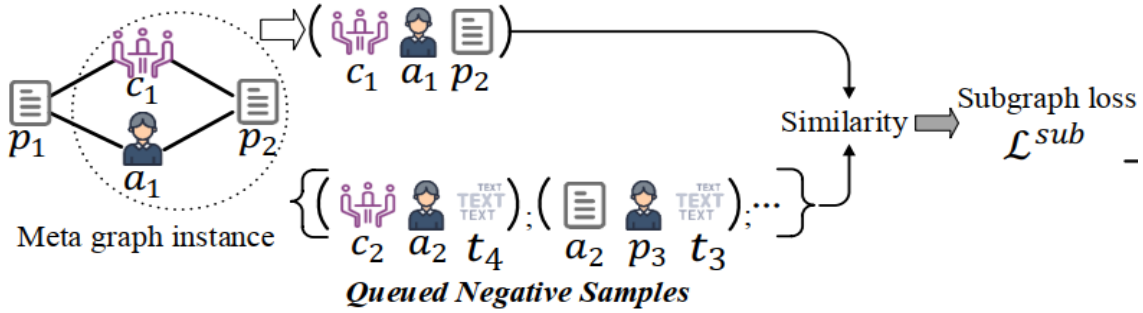
<author1, “writes”, paper1>

<author1, “joins”, conference>

<author1, “writes”, paper2>

Relation-level Pre-training Task

Persevere the high-order semantic contexts on a heterogeneous graph



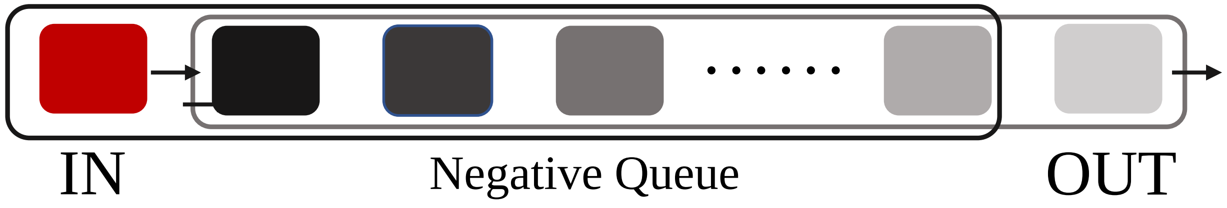
- Structural Positive Samples

  capture local connectivity and semantic contexts.

- Queued Negative Samples

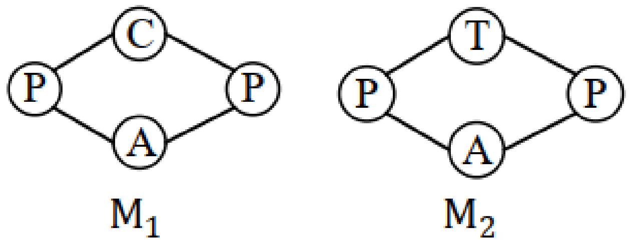
  negative samples from the latest positive sample,

  which leads the contrastive procedure more efficient

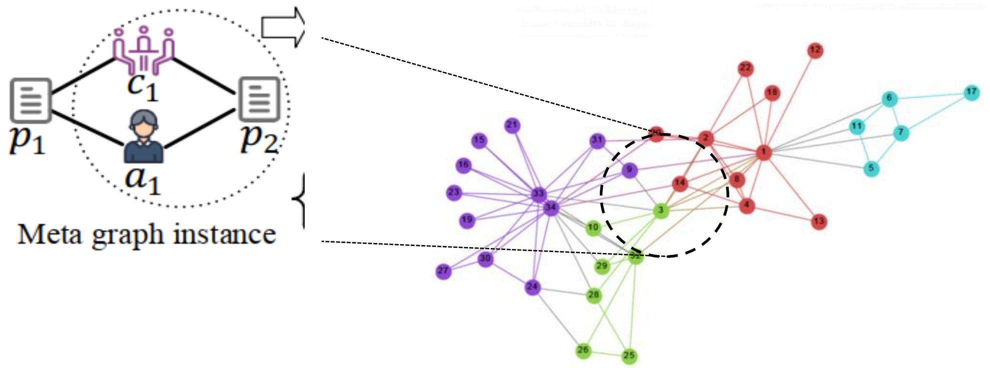


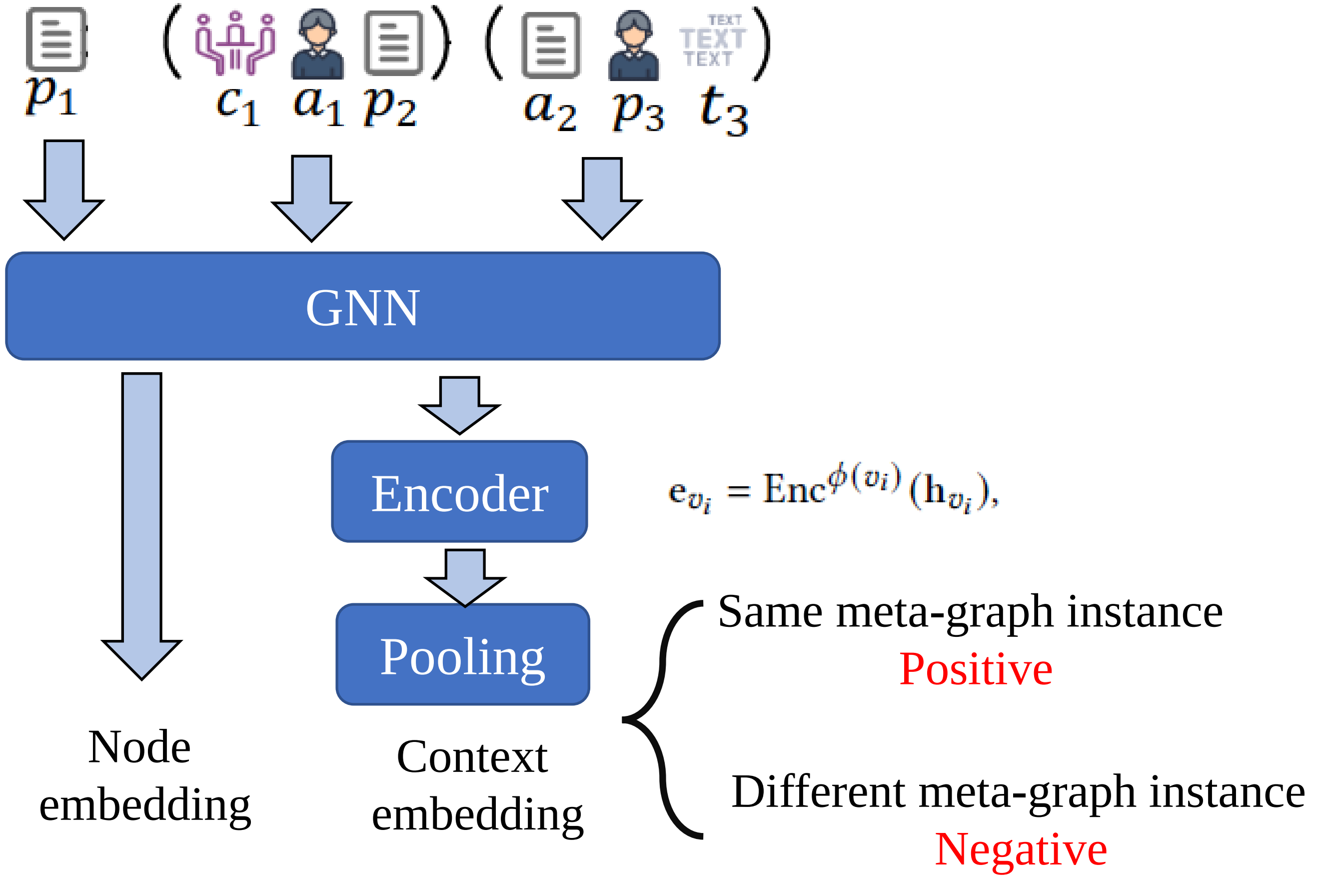
Subgraph-level pre-training task

sampling according to meta-graph



center node with **context nodes**





CONTENTS

Datasets and statistics

Dataset	#Node type	#Nodes	#Edge type	#Edges
DBLP	Author (A) Paper (P) Term (T) Venue (V)	4,057 16,670 13,420 40	P-A P-V P-T	19,645 16,670 133,039
Yelp	Business (B) Location (L) Star (S) Term (T)	7,474 65 9 36,412	B-T B-L B-S T-T	132,928 7,474 7,474 360,676
Aminer	Paper (P) Author (A) Conference (C) Terms (T)	614,209 737,621 842 80,589	P-A P-C P-P P-T	2,311,822 764,246 4,665,400 7,722,124

Baselines

- GAE
- DGI
- No-Pretrain
- GPT-GNN
- EdgePred

Experiment results on Node classification and Link Prediction

Dataset	Link Type	No pre-train	GAE	EgePred	DGI	GPT-GNN	CPT-HG	Improv.
DBLP	Paper-Term	12.34 ± 1.43	12.51 ± 0.71	12.61 ± 0.44	12.47 ± 0.68	12.71 ± 0.35	**13.06** ± 0.42	2.75%
Yelp	Business-Location	45.83 ± 0.42	45.92 ± 0.52	46.10 ± 0.31	45.57 ± 0.64	46.04 ± 0.75	**47.04** ± 0.71	2.03%
Aminer	Paper-Conference	39.23 ± 1.75	40.31 ± 0.78	39.86 ± 1.17	40.74 ± 1.35	41.37 ± 0.76	**42.17** ± 1.23	1.93%
	Paper-Author	5.63 ± 0.73	5.71 ± 0.41	5.62 ± 0.87	5.84 ± 0.52	6.02 ± 0.45	**6.43** ± 0.54	6.81%

Table 2: Experiment results (MRR ± std) in link prediction task on the three datasets. The best method is bolded, and the second best is underlined.

Downstream Task	Link Prediction			
Dataset	DBLP	Yelp	Aminer	
Link/Labeled Node Type	Paper-Term	Business-Location	Paper-Conference	Paper-Author
No pre-train	12.34 ± 1.43	45.83 ± 0.42	39.23 ± 1.75	5.63 ± 0.73
CPT-HG$_{sub}$	12.65 ± 0.42	**47.15** ± 0.44	41.54 ± 0.33	6.04 ± 0.51
CPT-HG$_{rel}$	12.79 ± 0.56	46.74 ± 0.65	41.75 ± 0.65	6.24 ± 0.15
CPT-HG	**13.06** ± 0.42	47.04 ± 0.71	**42.17** ± 1.23	**6.43** ± 0.54

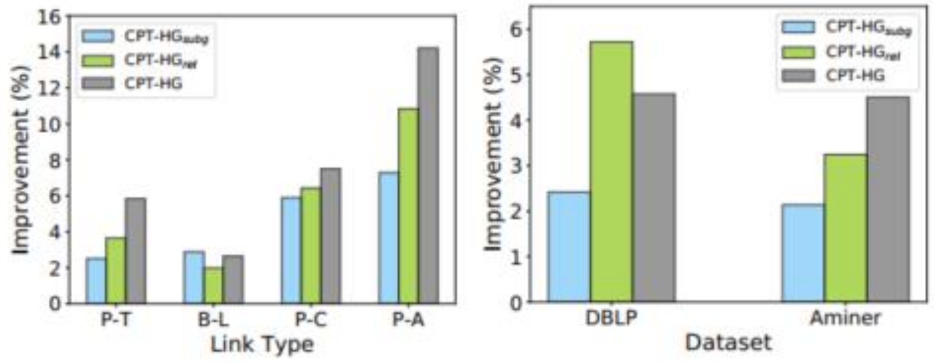
Table 4: Analysis of different ablated models in various downstream tasks.

- The improvements suggest that the contrastive pre-training on heterogeneous graphs is capable of learning transferable knowledge
- Both self-supervised tasks in GPT-GNN help the pre-training framework
  - Relation-level contrastive task
  - Subgraph-level contrastive task

Ablation study

Downstream Task	Link Prediction				Node Classification	
Dataset	DBLP	Yelp	Aminer		DBLP	Aminer
Link/Labeled Node Type	Paper-Term	Business-Location	Paper-Conference	Paper-Author	Paper	Author
No pre-train	12.34 ± 1.43	45.83 ± 0.42	39.23 ± 1.75	5.63 ± 0.73	87.45 ± 0.43	92.17 ± 0.56
CPT-HG$_{sub}$	12.65 ± 0.42	**47.15** ± 0.44	41.54 ± 0.33	6.04 ± 0.51	89.57 ± 0.61	94.14 ± 0.54
CPT-HG$_{rel}$	12.79 ± 0.56	46.74 ± 0.65	41.75 ± 0.65	6.24 ± 0.15	**92.45** ± 0.54	95.16 ± 0.32
CPT-HG	**13.06** ± 0.42	47.04 ± 0.71	**42.17** ± 1.23	**6.43** ± 0.54	91.45 ± 0.54	**96.32** ± 0.43

Table 4: Analysis of different ablated models in various downstream tasks.



(a) Link Prediction. (b) Node Classification.

Figure 3: Improvement of different ablated models over no pre-train baseline.

- The semantic relations encoded in CPT-HGrel seem more important for node representations in heterogeneous graph
- CPT-HGsub significantly improves the link prediction performance by focusing on modeling subgraph structures in a graph.

CONTENTS

Conclusion

- **CPT-HG**, which is a pre-training framework, enables the GNN to capture heterogeneous semantics and structural properties
- Extensive experiments three real-world heterogeneous graphs