A Comprehensive Study on Text-attributed Graphs: Benchmarking and Rethinking

Hao Yan*, **Chaozhuo Li***, Ruosong Long, Chao Yan, Jianan Zhao, Wenwen Zhuang, Jun Yin, Peiyan Zhang, Weihao Han, Hao Sun, Weiwei Deng, Qi Zhang, Lichao Sun, Xing Xie, Senzhang Wang

Central South University, Microsoft Research Asia, University of Birmingham,

Peking University, Hong Kong University of Science and Technology,

Université de Montréal, Microsoft, Lehigh University

NeurIPS 2023 Datasets and Benchmark track

CONTENT

BACKGROUND

- Graphs are ubiquitous in real word

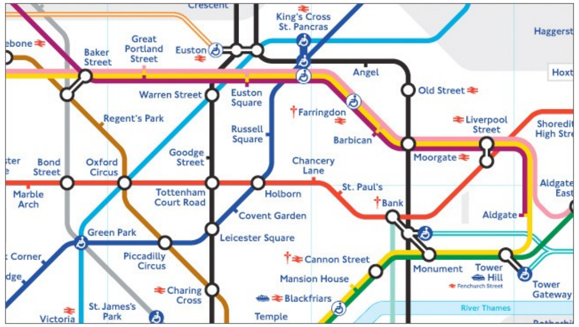


Image credit: visitlondon.com

- Underground Networks

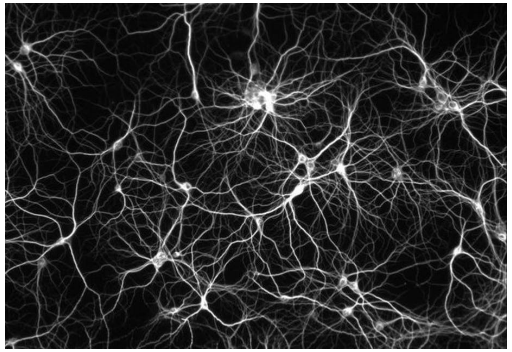


Image credit: The Conversation

- Networks of Neurons

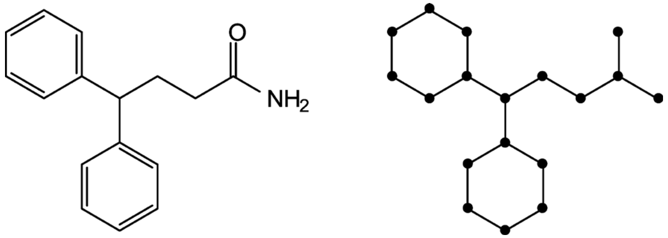


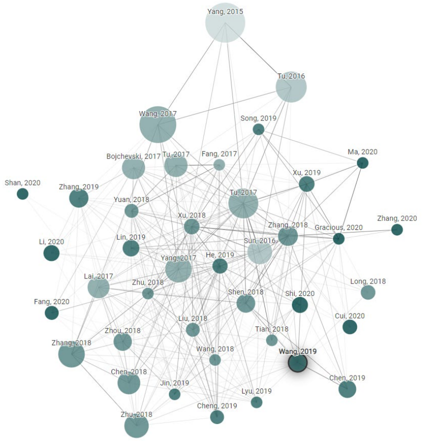
Image credit: MDPI

- Molecules

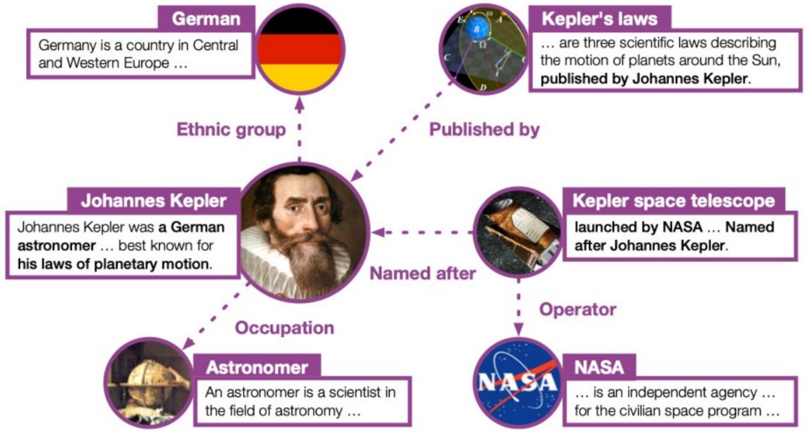
- Text-attributed graphs (TAGs): Nodes are often associated with text attributes

Image credit: Medium

- Social Networks



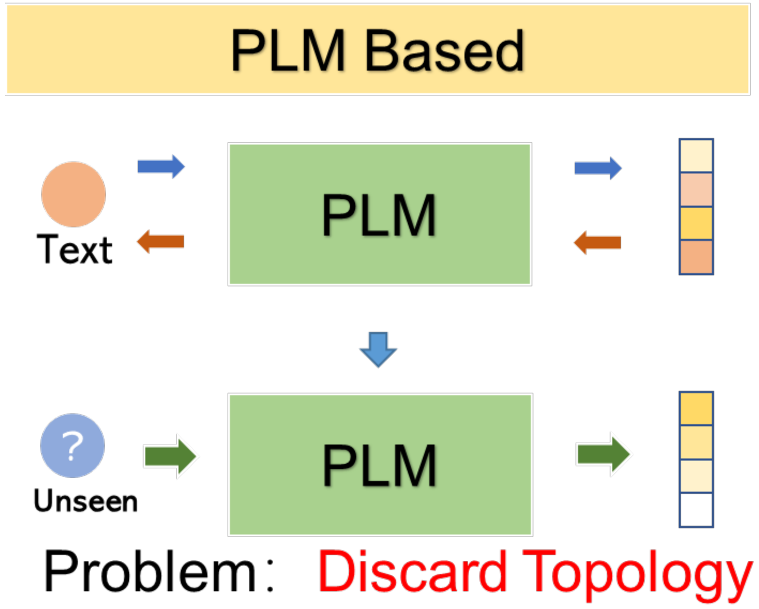
- Citation Networks



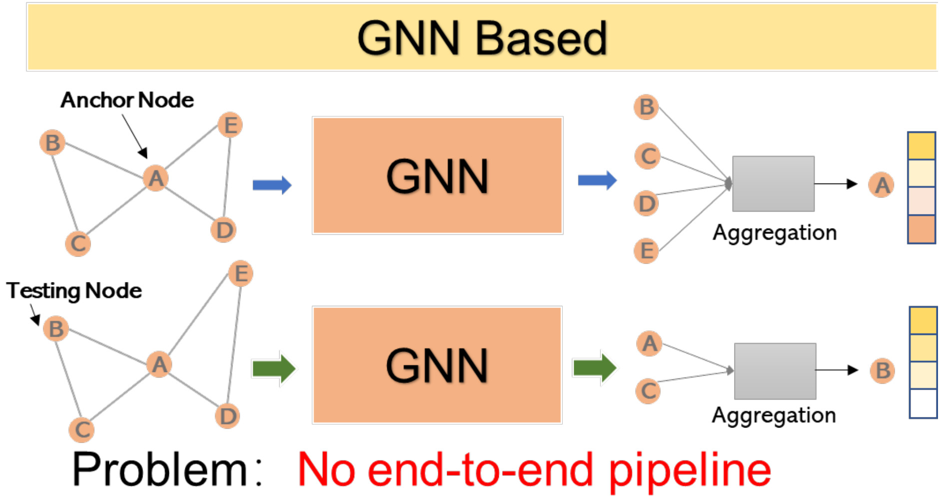
- Knowledge Graph

BACKGROUND

- Representation Learning on TAGs
  - Pre-trained Language Model (PLM) based
    - PLMs (Encoder-only) : BERT, ELECTRA, DeBERTa, RoBERTa.
    - Problem: The knowledge of topology is largely discarded



  - Graph Neural Network (GNN) based
    - GNNs (Message-Passing): GCN, GraphSAGE, GAT.
    - Problem: Modeling of node attributes and graph topology is disconnected

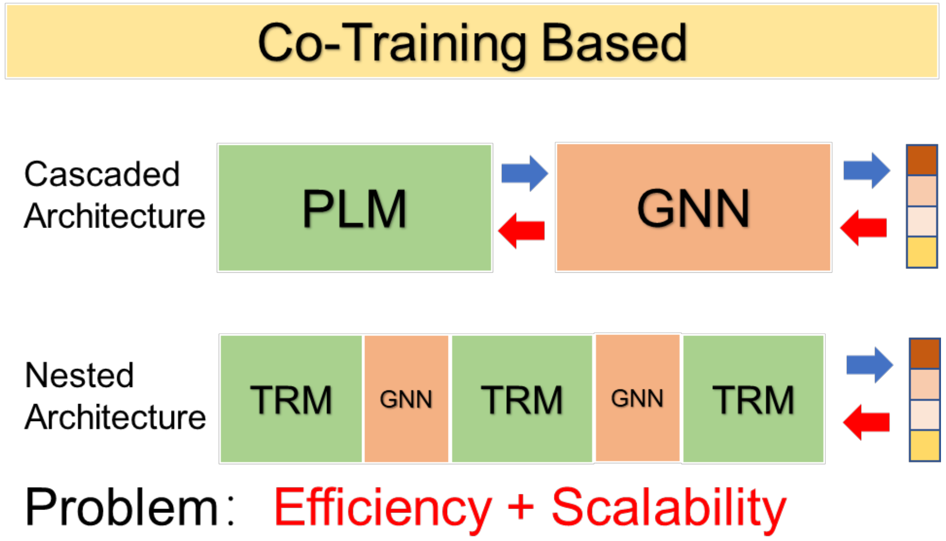


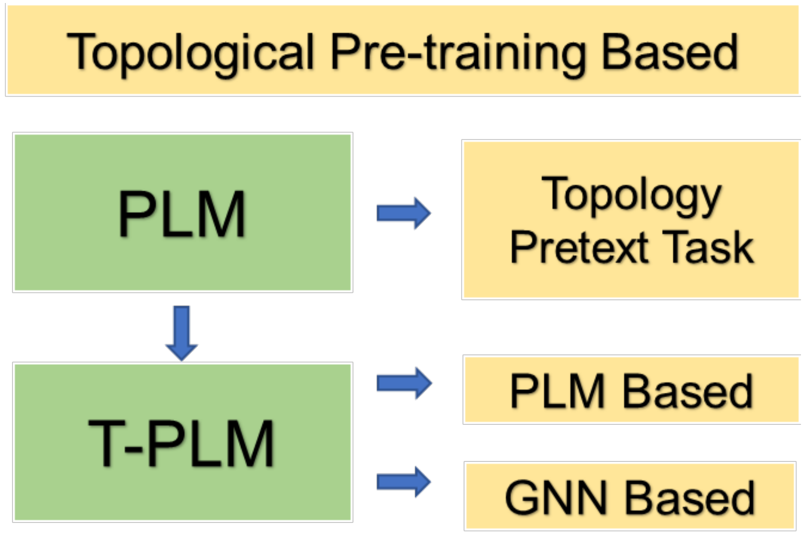
BACKGROUND

- Representation Learning on TAGs
  - Co-Training based
    - An end-to-end training paradigm to jointly model the node attributes and graph topology

$$f_{\Theta}(A,T) = GNN_{\omega}\left(A, LM_{\psi}(T)\right), \Theta = \{\omega, \psi\}$$

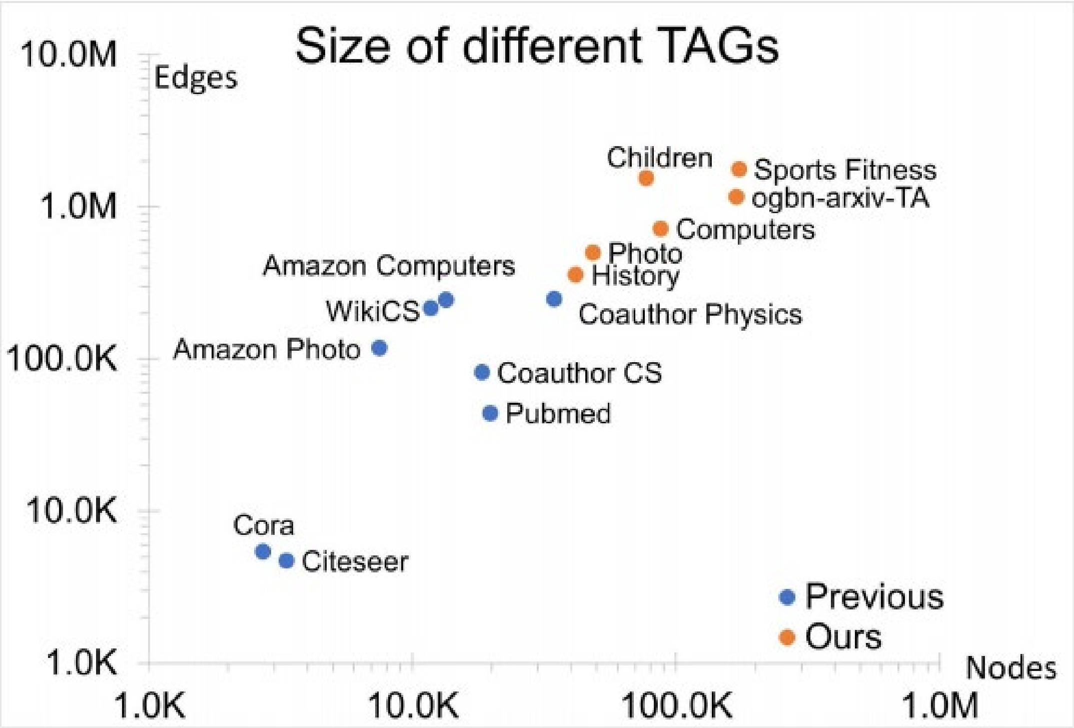
    - Problem: Scalability issues
  - Topological Pre-training based
    - Can we pre-train the LMs to understand the graph topology ?
    - How to design the pre-training tasks ?





DATASET

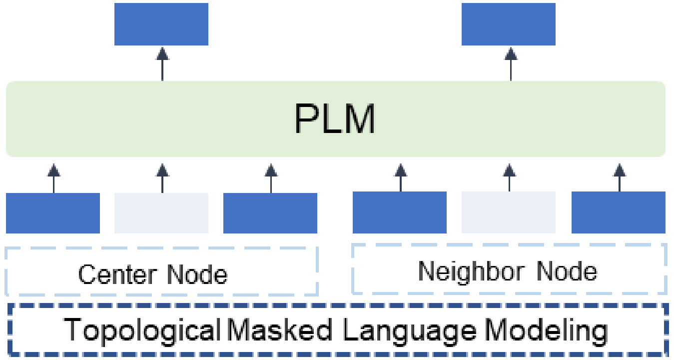
- Previous Datasets
  - Lack the availability of raw textual information
  - Overlook the exploration of text attribute modeling's impact on GNNs
  - These datasets are predominantly small in scale



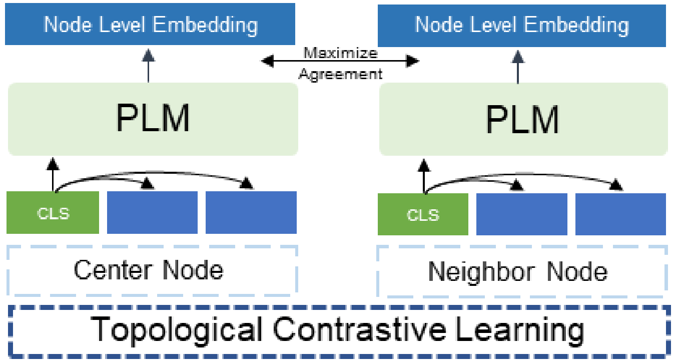
- CS-TAG
  - Keeping raw textual information
  - Modeling text-attributes: different PLMs
  - Scale larger

METHODS

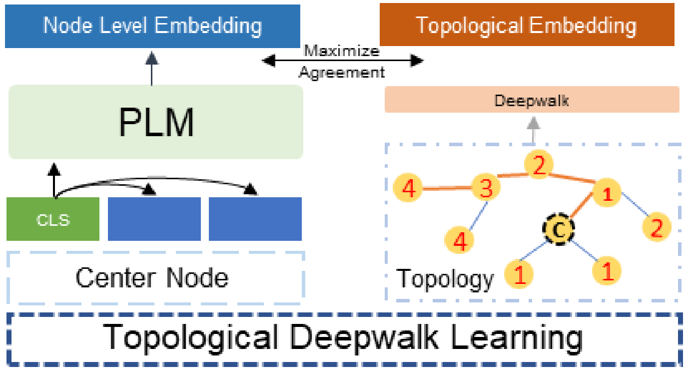
- Topological Masked Language Model (TMLM)
  - The first-order topological information in the token-level



- Topological Contrastive Learning (TCL)
  - The first-order topological information in the node-level



- Topological Deepwalk Learning (TDK)
  - The higher-order topological information in the node-level



EXPERIMENT

- GNNs with different PLM's features

Scale	PLMs	Arxiv GCN	GAT	SAGE	RevGAT	NFormer	GIN	JKNet	APPNP	MoNet	MLP
Small	BERT-Tiny	72.03	72.25	72.35	**72.52**	71.91	68.42	69.50	71.63	45.13	57.22
	ELECTRA	68.45	70.97	69.63	**71.12**	69.45	58.09	62.87	59.55	36.65	36.58
	DistilBERT	73.39	73.48	74.48	**74.68**	**73.56**	72.30	**71.44**	74.01	**50.51**	68.11
Base	ELECTRA	70.81	71.67	70.82	**71.96**	70.43	64.88	63.41	65.62	38.91	48.56
	BERT	73.30	73.40	74.14	**74.59**	72.80	71.94	70.08	73.90	46.90	67.35
	RoBERTa	73.56	73.38	74.52	**74.82**	73.12	72.63	69.40	74.01	44.53	69.31
	DeBERTa	68.15	66.56	67.58	**68.26**	67.11	62.05	44.16	52.37	29.67	47.07
Large	ELECTRA	70.44	71.01	70.72	**72.56**	70.04	64.47	58.34	64.52	37.26	47.72
	BERT	73.25	73.37	74.15	**74.68**	73.12	71.88	68.70	73.53	43.31	66.85
	RoBERTa	**73.95**	**73.72**	**74.64**	**74.99**	73.12	**73.10**	68.10	**74.17**	44.01	**69.51**
	DeBERTa	72.57	71.50	73.22	**73.59**	71.88	71.25	54.41	69.28	33.53	66.28
	Diff	5.80	7.16	7.06	7.54	6.45	15.01	27.28	21.80	20.84	32.93

EXPERIMENT

- Analyzing the performance of the Co-Training paradigm

Way	PLM-Based		GNN-Based				Co-Training Based	
	Tiny	Base	T-GCN	B-GCN	T-SAGE	B-SAGE	GCN(T)	SAGE(T)
Arxiv	70.83	72.96	72.03	73.30	72.35	**74.14**	69.22↓	73.57↑
Children	49.85	**59.91**	57.07	58.11	57.57	58.74	54.75↓	59.70↑
History	83.06	**86.09**	84.52	85.04	84.79	85.12	83.52↓	85.09↑
Photo	73.75	77.53	82.42	82.70	83.25	83.27	83.32↑	**86.64**↑
Computers	58.32	60.40	87.43	87.86	87.90	**88.30**	83.93↓	86.04↓
Sports	81.47	86.02	84.93	86.16	87.06	**87.34**	85.06↑	85.87↓

Datasets	BERT-Tiny						BERT-Base					
	Co-Training			TCL			Co-Training			TCL		
	Acc	Memory	Time	Acc	Memory	Time	Acc	Memory	Time	Acc	Memory	Time
Arxiv	73.57	76.27%	44.0	71.55	27.59%	7.0	-	OOM	-	**74.87**	70.73%	130
Children	59.70	97.28%	15.5	54.11	19.76%	2.0	-	OOM	-	**60.73**	80.99%	30
History	85.09	85.74%	5.7	86.06	14.69%	1.3	-	OOM	-	**86.80**	98.73%	18
Photo	**86.64**	97.83%	14.6	73.86	22.75%	3.1	-	OOM	-	82.85	70.65%	120
# Average	76.25	89.38%	19.95	71.40	21.20%	3.4	-	OOM	-	**76.31**	80.28%	74.50

EXPERIMENT

- Node classification results

Scale	Model	Arxiv						History					
		PLM	GNNs	TMLM	TDK	TCL	TMDC	PLM	GNNs	TMLM	TDK	TCL	TMDC
Small	BERT-Tiny	70.83	**72.52**	70.83	71.50	71.55	71.17	83.06	85.03	85.76	85.79	86.06	**86.88**
	ELECTRA	71.26	71.12	72.65	72.83	73.06	**73.71**	84.18	83.11	84.54	84.42	84.57	**85.18**
	DistilBERT	72.50	74.68	73.53	74.38	74.89	**75.50**	85.81	85.67	85.76	86.29	86.28	**86.88**
Base	ELECTRA	72.67	71.96	73.51	74.33	74.26	**75.56**	85.64	83.79	85.77	85.88	**86.62**	86.41
	BERT	72.96	74.59	73.97	74.23	74.87	**76.11**	86.09	85.28	86.24	86.46	86.80	**86.82**
	RoBERTa	73.10	74.82	74.25	74.57	75.37	**75.97**	85.85	85.69	86.19	86.32	86.95	**86.96**
	DeBERTa	73.82	68.26	74.26	75.01	75.15	**75.99**	86.16	82.31	86.00	86.46	**87.01**	86.94
Large	ELECTRA	72.42	72.56	74.76	73.82	74.17	**75.58**	86.13	83.56	86.39	86.49	**86.82**	86.28
	BERT	73.24	74.68	75.01	74.31	75.15	**75.75**	86.24	85.15	86.47	86.73	86.93	**86.94**
	RoBERTa	73.83	74.99	75.18	74.58	75.48	**75.73**	86.41	85.23	86.72	86.75	87.11	**87.22**
	DeBERTa	74.57	73.59	75.92	75.20	75.58	**76.20**	87.00	84.89	87.11	87.26	87.30	**87.32**
Scale	**Model**	**Children**						**Photo**					
		PLM	GNNs	TMLM	TDK	TCL	TMDC	PLM	GNNs	TMLM	TDK	TCL	TMDC
Small	BERT-Tiny	49.85	**57.86**	54.27	53.43	54.11	54.66	73.75	**84.12**	74.30	73.99	73.86	74.92
	ELECTRA	57.03	56.42	57.35	56.92	56.88	**58.55**	76.58	**83.12**	76.09	76.89	77.74	77.83
	DistilBERT	59.90	59.33	60.03	60.23	60.60	**61.38**	77.51	**84.34**	77.81	79.69	81.85	82.52
Base	ELECTRA	59.09	56.42	59.93	60.27	60.21	**60.83**	77.84	**82.98**	78.27	80.18	81.47	82.82
	BERT	59.91	58.74	60.34	60.43	60.73	**61.43**	77.53	**84.46**	78.54	81.04	82.85	84.09
	RoBERTa	59.80	59.01	60.19	60.71	61.47	**61.83**	78.11	**84.59**	78.33	81.26	82.47	83.04
	DeBERTa	60.26	50.72	60.73	61.39	61.92	**62.20**	78.37	81.44	79.27	81.34	83.07	**83.80**
Large	ELECTRA	58.28	56.59	60.51	59.31	59.29	**61.31**	77.25	**83.00**	79.21	78.44	79.56	81.32
	BERT	60.65	58.90	60.84	61.15	61.50	**62.06**	77.72	**84.21**	78.95	79.26	80.74	81.14
	RoBERTa	60.93	59.26	62.11	61.95	62.06	**63.24**	79.60	**85.12**	80.32	80.82	81.47	82.55
	DeBERTa	61.61	56.34	61.91	**62.51**	62.37	62.46	79.63	82.55	80.45	81.33	82.33	**82.70**

CONCLUSION

- First comprehensive benchmark on TAGs: CS-TAG.
  - Collect and provide text-attributed graph datasets.
  - Comprehensive evaluation of different learning paradigms.
- Topological Pre-training of Language Models
  - TMLM, TDK, TCL, TMDC

NEURAL INFORMATION PROCESSING SYSTEMS

THANKS