NEURAL INFORMATION PROCESSING SYSTEMS

TEG-DB: A Comprehensive Dataset and Benchmark of Textual-Edge Graphs

Zhuofeng Li*△, Zixing Gou*▽, Xiangnan Zhang◇, Zhongyuan Liu◦, Sirui Li†, Yuntong Hu†, Chen Ling†, Zheng Zhang†, Liang Zhao†

△Shanghai University, ▽Shandong University, ◇Johns Hopkins University, ◦China University of Petroleum (East China), †Emory University

* These two authors contribute equally to this paper.

CONTENT

BACKGROUND

- Graphs in the world are ubiquitous, diverse and entangled.

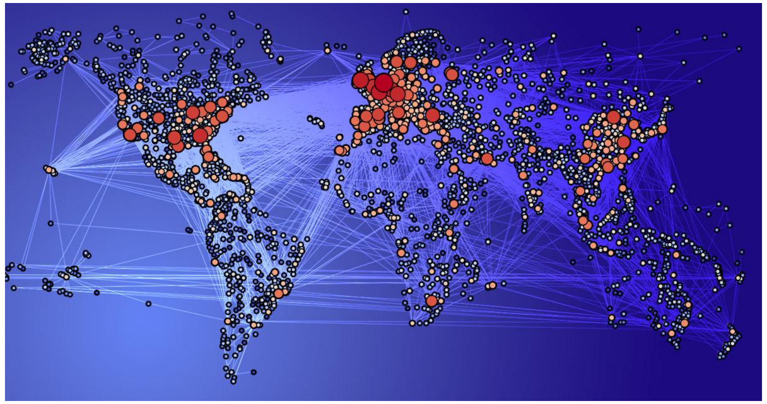


Image credit : Télécom Paris

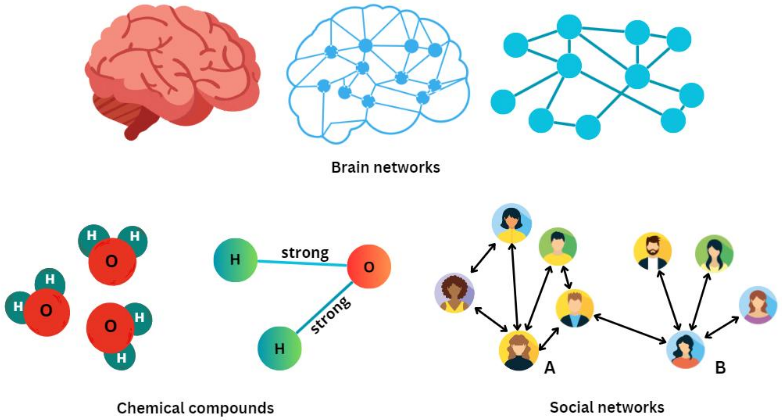


Image credit : Medium

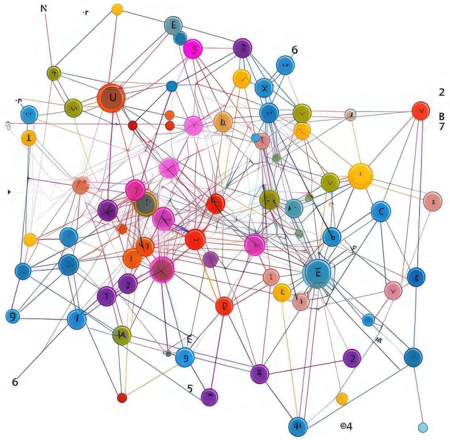
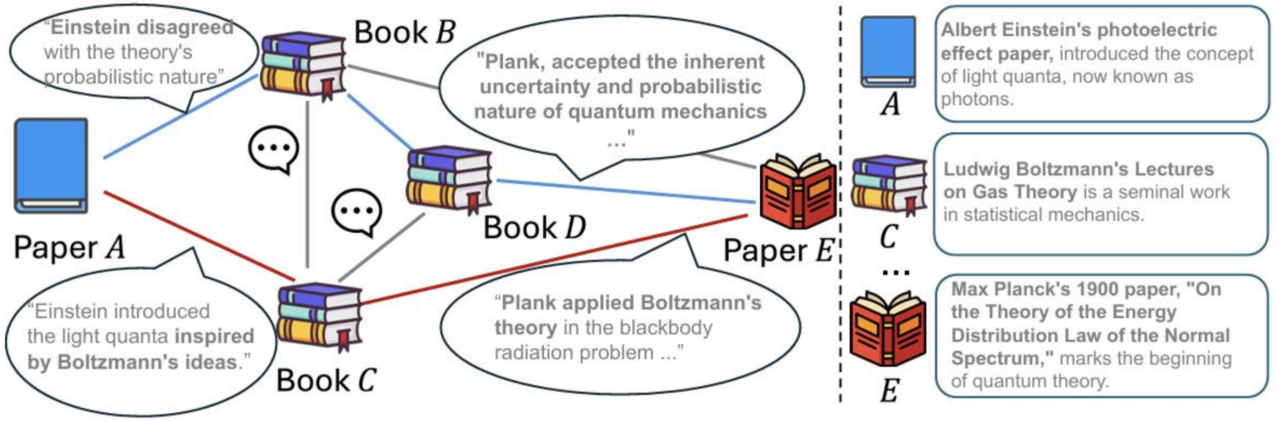


Image credit : Medium

- Ubiquitous
- Diverse
- Entangled

- Textual-Edge Graphs (TEGs): Rich textual descriptions on nodes and edges.



Textual-edge Graph Example

Scientific articles in quantum theory linked by citations.

BACKGROUND

- Representation Learning on TEGs
  - Pre-trained Language Model (PLM) based methods
    - LLMs: Llama, PaLM, GPT
    - Problem: Ignore the topology among graphs

$$\boldsymbol{h}_u^{(k+1)} = \mathrm{MLP}_{\boldsymbol{\psi}}^{(k)}\left(\boldsymbol{h}_u^{(k)}\right)$$

$$\boldsymbol{h}_u^{(0)} = \mathrm{PLM}(T_u) + \sum_{v \in \mathcal{N}(u)} \mathrm{PLM}(T_{e_{v,u}})$$

  - Graph Neural Network (GNN) based methods
    - GNNs: GCN, GAT, GraphSAGE, GIN, RevGAT
    - Problem: Fall short of fully capture semantic information

$$\boldsymbol{h}_u^{(k+1)} = \mathrm{UPDATE}_{\boldsymbol{\omega}}^{(k)}\left(\boldsymbol{h}_u^{(k)}, \mathrm{AGGREGATE}_{\boldsymbol{\omega}}^{(k)}\left(\left\{\boldsymbol{h}_v^{(k)}, \boldsymbol{e}_{v,u}, v \in \mathcal{N}(u)\right\}\right)\right)$$

BACKGROUND

- Representation Learning on TEGs
  - LLM as predictor
    - Problems: Information loss and limitations in efficiency

$$A = f\{\mathcal{G}, Q\}$$

  - Benchmarks for existing text-attributed graphs
    - First stage datasets: mag, ogbn-arxiv
    - Second stage: CS-TAG
    - Problems:
      1. Include texts only on nodes
      2. Lack coverage across diverse domains and tasks
      3. Lack of uniformity in representation formats

DATASETS

- Previous Datasets
  - Overlook the text information from the edge
  - Lack a standardized data format
  - TEG datasets are inadequate

- Improved Datasets —— TEG
  - Rich textual descriptions on both nodes and edges
  - Cover a wide range of domains and sizes.
  - Unified format

	Dataset	Nodes	Edges	Nodes-Class	Graph Domain	Size	Nodes-text	Edges-text	Node Classification	Link Prediction
Previous	Twitch Social Network [35]	7,126	88,617	2	Social Networks	Small	✗	✗	✓	✗
	Facebook Page-Page Network [36]	22,470	171,002	4	Social Networks	Small	✗	✗	✓	✗
	ogbn-arxiv [13]	169,343	1,166,243	40	Academic	Medium	✓	✗	✓	✗
	Citeseer [37]	3,327	4,732	6	Academic	Small	✗	✗	✓	✗
	Pubmed [37]	19,717	44,338	3	Academic	Small	✗	✗	✓	✗
	Cora [28]	2,708	5,429	7	Academic	Small	✗	✗	✓	✗
	CitationV8 [46]	1,106,759	6,120,897	-	Academic	Large	✓	✗	✗	✓
	GoodReads [46]	676,084	8,582,324	11	Book Recommendation	Large	✓	✗	✗	✓
	Sports-Fitness [46]	173,055	1,773,500	13	E-commerce	Medium	✓	✗	✓	✗
	Ele-Photo [46]	48,362	500,928	12	E-commerce	Small	✓	✗	✓	✗
	Books-History [46]	41,551	358,574	12	E-commerce	Small	✓	✗	✓	✗
	Books-Children [46]	76,875	1,554,578	24	E-commerce	Small	✓	✗	✓	✗
	ogbn-arxiv-TA [46]	169,343	1,166,243	40	Academic	Medium	✓	✗	✓	✗
Ours	Goodreads-History	540,807	2,368,539	11	Book Recommendation	Large	✓	✓	✓	✓
	Goodreads-Crime	422,653	2,068,223	11	Book Recommendation	Large	✓	✓	✓	✓
	Goodreads-Children	216,624	858,586	11	Book Recommendation	Large	✓	✓	✓	✓
	Goodreads-Comics	148,669	631,649	11	Book Recommendation	Medium	✓	✓	✓	✓
	Amazon-Movie	137,411	2,724,028	399	E-commerce	Medium	✓	✓	✓	✓
	Amazon-Apps	31,949	62,036	62	E-commerce	Small	✓	✓	✓	✓
	Reddit	478,022	676,684	3	Social Networks	Large	✓	✓	✓	✓
	Twitter	18,761	23,764	-	Social Networks	Small	✓	✓	✗	✓
	Citation	4,972,456	5,970,965	24	Academic	Large	✓	✓	✓	✓

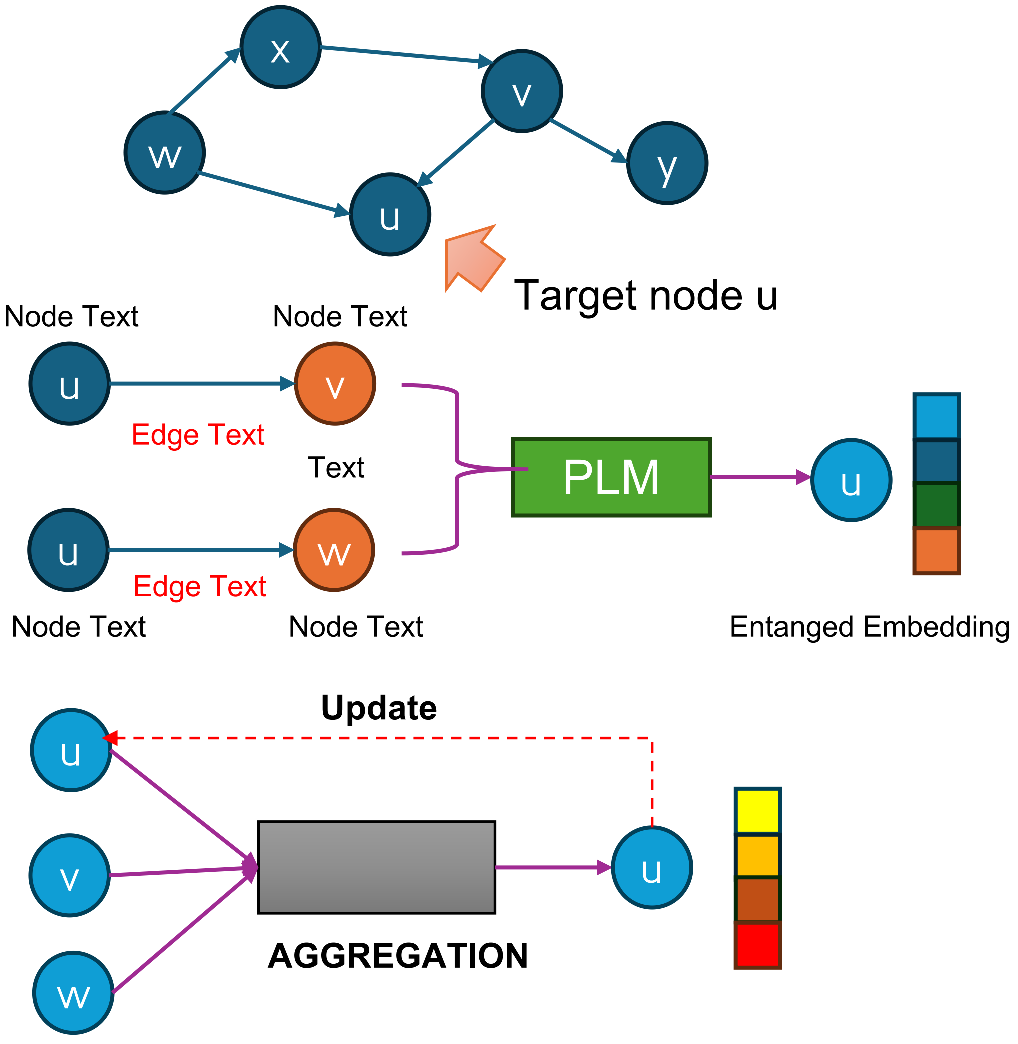
METHODS

- Entangled GNN-based Paradigm
  - Entangled edge-text encoding with node-aware tokens
    - Initialize node representation using PLM

$$h_u^0 = PLM(T_u, \{T_v, T_{e_{v,u}}, v \in N(u)\})$$

    - Message Update

$$h_u^{(k+1)} = UPDATE_\omega^{(k)}(h_u^{(k)}, AGGREGATE_\omega^{(k)}(\{h_v^{(k)}, v \in N(u)\}))$$



Node u Representation via Entangled Edge-Text Encoding

EXPERIMENTS

- Link prediction among GNN-based methods

Methods	Children										Crime									
	Entangled-GPT		GPT-3.5-TURBO		BERT-Large		BERT		None		Entangled-GPT		GPT-3.5-TURBO		BERT-Large		BERT		None	
	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1
MLP	0.9146	0.8459	0.8952	0.8198	0.8948	0.8193	0.8947	0.8192	0.8929	0.8181	0.9030	0.8429	0.8911	0.8144	0.8909	0.8145	0.8920	0.8153	0.8913	0.8149
GraphSAGE	**0.9744**	0.9011	**0.9520**	0.8866	0.9493	0.8821	0.9503	0.8848	0.9400	0.8736	0.9331	0.8629	0.9241	0.8541	0.9537	0.8887	0.9529	0.8868	0.9053	0.8320
General GNN	0.9653	0.9015	0.9519	0.8907	**0.9521**	**0.8921**	**0.9540**	**0.8953**	0.9356	0.8735	**0.9356**	**0.8792**	**0.9325**	**0.8625**	**0.9568**	**0.8957**	0.9257	0.8526	0.9117	0.8426
GINE	0.9558	**0.9132**	0.9518	**0.8939**	0.9463	0.8878	0.9491	0.8914	**0.9389**	0.8748	0.9324	0.8589	0.9125	0.8429	0.9517	0.8878	**0.9538**	**0.8928**	**0.9132**	**0.8448**
EdgeGNN	0.9604	0.9055	0.9487	0.8851	0.9488	0.8884	0.9504	0.8891	0.9352	**0.8765**	0.9309	0.8575	0.9104	0.8410	0.9545	0.8914	0.9535	0.8897	0.9036	0.8345
GraphTransformer	0.9625	0.8950	0.9487	0.8751	0.9441	0.8742	0.9431	0.8763	0.9241	0.8333	0.9123	0.8592	0.9078	0.8309	0.9465	0.8769	0.9479	0.8817	0.8985	0.8256

Methods	Amazon-Apps										Amazon-Movie									
	Entangled-GPT		GPT-3.5-TURBO		BERT-Large		BERT		None		Entangled-GPT		GPT-3.5-TURBO		BERT-Large		BERT		None	
	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1
MLP	0.8950	0.7980	0.8642	0.7752	0.8639	0.7698	0.8634	0.7698	0.8655	0.7738	0.8509	0.7490	0.8227	0.7269	0.8349	0.7553	0.8349	0.7555	0.8205	0.7317
GraphSAGE	0.8911	0.8073	0.8662	0.7853	**0.8813**	0.7971	0.8783	0.8015	0.8634	0.7366	0.8725	0.7911	0.8500	0.7665	0.9067	0.8298	0.9178	0.8426	0.8507	0.7591
General GNN	**0.8956**	0.8340	**0.8810**	0.8178	0.8768	0.8131	0.8757	0.8090	**0.8680**	**0.8129**	**0.8849**	0.8134	**0.8659**	**0.7928**	**0.9206**	**0.8485**	0.8937	**0.8483**	**0.8617**	**0.7918**
GINE	0.8875	0.8179	0.8559	0.8099	0.8680	0.8092	0.8555	0.8123	0.8671	0.8065	0.8712	**0.8154**	0.8603	0.7911	0.9187	0.8454	0.9165	0.8456	0.8591	0.7879
EdgeGNN	0.8956	**0.8403**	0.8720	**0.8180**	**0.8813**	**0.8153**	**0.8804**	**0.8184**	0.8520	0.8043	0.8708	0.8035	0.8565	0.7842	0.9171	0.8436	**0.9181**	0.8468	0.8552	0.7837
GraphTransformer	0.8634	0.7820	0.8395	0.7647	0.8748	0.7926	0.8736	0.7846	0.8469	0.7329	0.8537	0.7698	0.8339	0.7453	0.9035	0.8196	0.9044	0.8185	0.8393	0.7550

Methods	Citation										Twitter									
	Entangled-GPT		GPT-3.5-TURBO		BERT-Large		BERT		None		Entangled-GPT		GPT-3.5-TURBO		BERT-Large		BERT		None	
	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1
MLP	0.9251	0.8679	0.9170	0.8598	0.9173	0.8561	0.8935	0.8613	0.8857	0.8015	0.7085	0.5669	0.6991	0.5430	0.8115	0.7898	0.8136	0.7148	0.7007	0.5430
GraphSAGE	0.9494	0.8972	0.9369	0.8758	**0.9457**	**0.8832**	0.9780	0.9300	0.8925	0.8345	0.6998	0.6486	0.6779	0.6193	0.8609	0.8177	0.8359	0.7964	0.5668	0.5940
General GNN	0.9470	0.8840	0.9258	0.8739	0.9281	0.8637	0.9327	0.8757	0.8984	0.8397	**0.8118**	**0.7247**	0.7888	0.7094	0.8531	0.7756	0.8062	0.6552	0.7017	**0.6163**
GINE	**0.9538**	**0.9085**	**0.9482**	**0.8939**	0.9443	0.8825	0.9736	0.9272	0.8744	0.8145	0.6835	0.6345	0.6696	0.6135	0.8306	0.7719	0.8738	0.7880	**0.7213**	0.6161
EdgeGNN	0.7382	0.5545	0.7136	0.5393	0.7132	0.5352	0.7401	0.6526	0.6965	0.5449	0.6940	0.6214	0.6854	0.6123	0.8290	0.6614	0.7513	0.6745	0.6124	0.5664
GraphTransformer	0.9536	0.8963	0.9350	0.8697	0.9439	0.8713	**0.9789**	**0.9320**	**0.9172**	**0.8441**	0.7030	0.6824	0.6859	0.6764	**0.8967**	**0.8223**	**0.8768**	**0.8165**	0.5908	0.5423

EXPERIMENTS

- Node Classification among GNN-based methods

Method	Children										Crime									
	Entangled-GPT		GPT-3.5-TURBO		BERT-Large		BERT		None		Entangled-GPT		GPT-3.5-TURBO		BERT-Large		BERT		None	
	AUC*	F1*	AUC*	F1*	AUC*	F1*	AUC*	F1*	AUC*	F1*	AUC*	F1*	AUC*	F1*	AUC*	F1*	AUC*	F1*	AUC*	F1*
MLP	0.8785	0.5904	0.8505	0.5663	0.8593	0.5810	0.8597	0.5749	0.8452	0.5811	0.9253	0.6842	0.9149	0.6615	0.9150	0.6619	0.9151	0.6602	0.9154	0.6624
GraphSAGE	**0.9569**	**0.8041**	0.9342	**0.7871**	**0.9162**	0.7497	**0.9152**	0.7440	**0.8713**	0.6227	0.9663	0.8325	**0.9549**	0.8189	0.9445	0.7832	0.9463	0.7848	0.9221	0.7048
General GNN	0.9534	0.7942	**0.9352**	0.7846	0.9161	**0.7502**	**0.9152**	0.7451	0.8681	0.6162	**0.9732**	**0.8437**	0.9546	**0.8200**	0.9446	**0.7854**	0.9456	**0.7888**	**0.9225**	**0.7262**
GINE	0.9529	0.7930	0.9324	0.7777	0.9154	0.7466	0.9137	**0.7552**	0.8523	0.6558	0.9636	0.8260	0.9504	0.8073	0.9410	0.7766	0.9429	0.7852	0.9155	0.7117
EdgeGNN	0.9542	0.7890	0.9338	0.7808	0.9128	0.7463	0.9121	0.7452	0.8583	0.6466	0.9581	0.8179	0.9490	0.8052	0.9400	0.7657	0.9405	0.7726	0.9187	0.6830
GraphTransformer	0.9525	0.7902	0.9340	0.7823	0.9137	0.7497	0.9150	0.7491	0.8517	**0.6565**	0.9613	0.8322	0.9505	0.8151	**0.9452**	0.7795	**0.9464**	0.7834	0.9220	0.6944

Method	Amazon-Apps										Amazon-Movie									
	Entangled-GPT		GPT-3.5-TURBO		BERT-Large		BERT		None		Entangled-GPT		GPT-3.5-TURBO		BERT-Large		BERT		None	
	AUC*	F1*	AUC*	F1*	AUC*	F1*	AUC*	F1*	AUC*	F1*	AUC*	F1*	AUC*	F1*	AUC*	F1*	AUC*	F1*	AUC*	F1*
MLP	0.7750	0.3429	0.7520	0.3204	0.8935	**0.4169**	0.8970	0.3107	0.7352	0.3067	0.9736	**0.5475**	0.9618	**0.5279**	0.9752	**0.5331**	0.9750	0.5173	0.9493	0.4625
GraphSAGE	**0.9439**	**0.4114**	**0.9274**	**0.3899**	**0.9226**	0.3794	**0.9229**	**0.3929**	**0.9161**	0.3348	0.9764	0.5325	0.9674	0.5165	**0.9773**	0.4919	**0.9771**	**0.5185**	0.9681	0.5096
General GNN	0.9138	0.3806	0.8947	0.3604	0.9171	0.3817	0.9223	0.3803	0.9151	**0.3932**	**0.9969**	0.5301	**0.9775**	0.5156	0.9768	0.4827	0.9768	0.5006	**0.9757**	0.5115
GINE	0.9356	0.3862	0.9170	0.3588	0.9170	0.2623	0.9185	0.3592	0.9028	0.3507	0.9732	0.4531	0.9507	0.4246	0.9758	0.4781	0.9759	0.5085	0.9168	0.4127
EdgeGNN	0.8857	0.3749	0.8764	0.3477	0.8639	0.2739	0.8800	0.3063	0.8568	0.2247	0.9483	0.5224	0.9360	0.5060	0.9372	0.4672	0.9263	0.4743	0.9492	0.4853
GraphTransformer	0.9400	0.3772	0.9195	0.3548	0.9217	0.3425	0.9225	0.3818	0.9155	0.3860	0.9910	0.5285	0.9763	0.5175	0.9764	0.4856	**0.9771**	0.5124	0.9756	**0.5126**

Method	Citation										Twitter									
	Entangled-GPT		GPT-3.5-TURBO		BERT-Large		BERT		None		Entangled-GPT		GPT-3.5-TURBO		BERT-Large		BERT		None	
	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1
MLP	0.7892	0.7879	0.7868	0.7859	0.7515	0.7471	0.8044	0.8032	**0.7493**	**0.7471**	0.8253	0.7549	0.8115	0.7261	0.8361	0.8193	0.8533	0.8329	0.8196	0.7383
GraphSAGE	0.7984	**0.8144**	0.7883	0.7874	0.7559	0.7525	0.8046	0.8060	0.7341	0.7308	0.8614	0.8055	0.8411	0.7903	0.8446	0.8305	0.8384	0.8247	0.8286	0.7802
General GNN	**0.8079**	0.8042	0.7906	0.7889	0.7546	0.7526	0.8057	0.8042	0.7361	0.7337	**0.8725**	**0.8574**	**0.8610**	0.8397	0.8368	0.8131	0.8609	0.8513	0.8401	0.8089
GINE	0.8055	0.8141	**0.7934**	**0.7925**	**0.7599**	**0.7574**	**0.8106**	**0.8100**	0.7316	0.7284	0.8649	0.8386	0.8438	0.8186	0.8401	0.8255	0.8460	0.8328	0.8254	0.7907
EdgeGNN	0.4261	0.3957	0.4140	0.3845	0.4082	0.3763	0.4200	0.3906	0.3935	0.3541	0.8714	0.8530	0.8551	**0.8442**	**0.8649**	**0.8574**	**0.8694**	**0.8607**	**0.8529**	**0.8431**
GraphTransformer	0.8022	0.7944	0.7903	0.7885	0.7531	0.7517	0.8070	0.8056	0.7369	0.7351	0.8720	0.8369	0.8563	0.8273	0.8342	0.8211	0.8402	0.8261	0.8197	0.7888

EXPERIMENTS

- Link Prediction among PLM-based methods

Methods	Goodreads-Children		Goodreads-Crime		Amazon-Apps		Amazon-Movie		Citation		Twitter	
	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1
GPT-3.5-TURBO	0.4770	0.1413	0.4507	0.1104	0.5000	**0.5200**	0.4843	0.1342	**0.8860**	**0.3514**	**0.4800**	0.3312
GPT-4	**0.8780**	**0.6090**	**0.8890**	**0.6040**	**0.6212**	0.1413	**0.5000**	**0.3000**	0.4735	0.3184	0.4300	**0.6144**

- Node Classification among PLM-based methods

Methods	Goodreads-Children		Goodreads-Crime		Amazon-Apps		Amazon-Movie		Citation	
	AUC*	F1*	AUC*	F1*	AUC*	F1*	AUC*	F1*	ACC	F1
GPT-3.5-TURBO	0.5200	0.0300	0.5400	0.0700	**0.5000**	**0.0100**	0.5159	0.0017	0.7098	0.3402
GPT-4	**0.6700**	**0.1800**	**0.6100**	**0.1400**	0.4995	0.0002	**0.5175**	**0.0029**	**0.8432**	**0.8450**

CONCLUSION

- **Introduction of TEG-DB**: The first TEG benchmark, designed to advance graph representation learning on TEGs by incorporating textual content on both nodes and edges, unlike traditional TAGs.
- **Comprehensive Dataset Collection**: Provides nine extensive textual-edge datasets to encourage collaboration between NLP and GNN communities.
- **Benchmark for Learning Approaches**: Offers an in-depth evaluation of various methods, highlighting their strengths and limitations.
- **Future Commitment**: Expand and develop research-oriented TEGs to support the field’s ongoing growth and innovation.

NEURAL INFORMATION PROCESSING SYSTEMS

THANKS