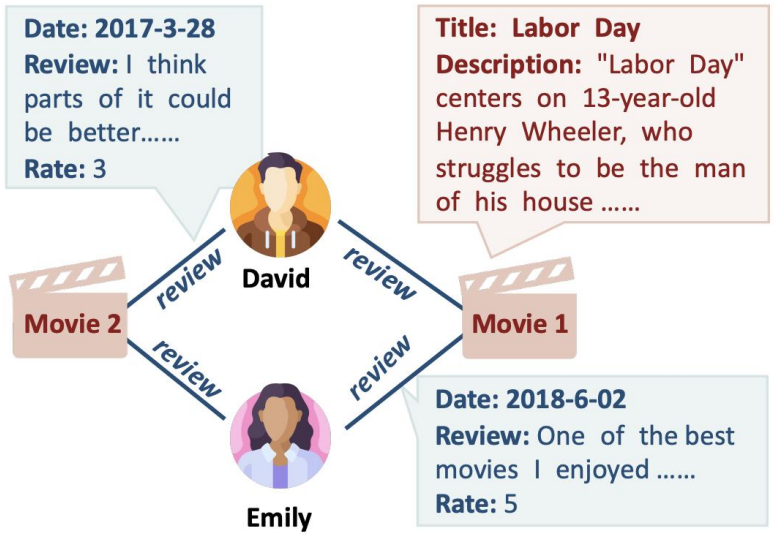
DTGB: A Comprehensive Benchmark for Dynamic Text-Attributed Graphs

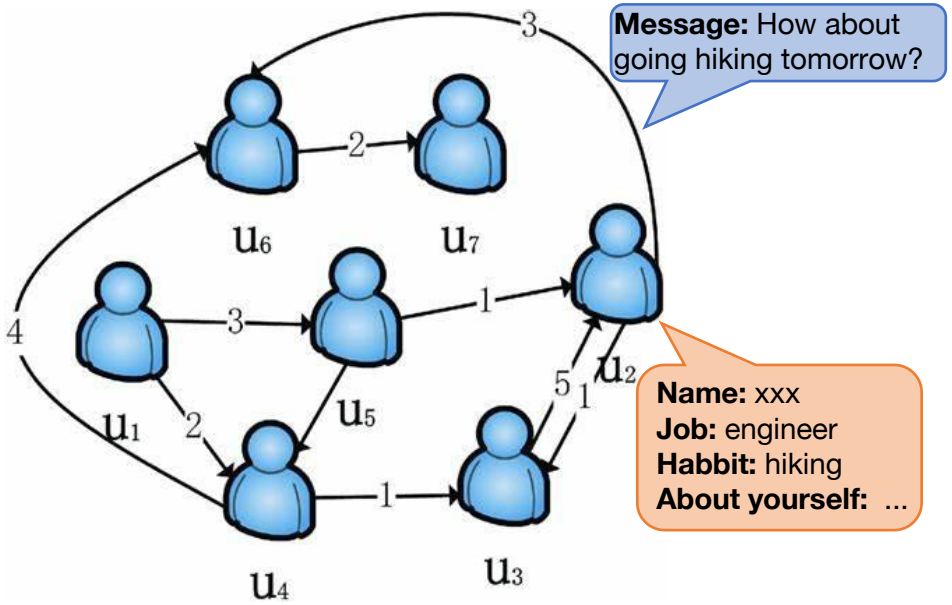
Jiasheng Zhang, Jialin Chen, Menglin Yang, Aosong Feng, Shuang Liang,
Jie Shao, and Rex Ying

NeurIPS 2024 Datasets and Benchmarks Track

Background



E-commerce Platform



Social Network

Temporal Knowledge Graph

Explainable Recommendation

Public Sentiment Detection

Political Relationship Analysis

Limitation of Existing Datasets

(1) Dynamic Graph Datasets

- MOOC [1], LastFM [2]

  Small scale (usually less than 10k nodes)
  Lack raw node and edge text descriptions
  Lack reasonable time segmentation (LastFM has 1.2 million edges and also 1.2 million different timestamps)

- TGB datasets [3]

  Large scale
  More reasonable time segmentation
  Lack raw node and edge text descriptions
  Lack auxiliary labels (e.g., edge categories)

[1] Predicting dynamic embedding trajectory in temporal interaction networks
[2] Towards Better Dynamic Graph Learning: New Architecture and Unified Library
[3] Temporal Graph Benchmark for Machine Learning on Temporal Graphs

Limitation of Existing Datasets

(1) Text-attributed Graph Datasets

- Cora [1], ogbn-arxiv [2]

  Small scale (usually less than 10k nodes)
  Lack raw node and edge text descriptions (only provide text-based features)
  Lack time information

- CS-TAG datasets [3]

  Large scale
  Node text descriptions
  Lack time information
  Lack edge text descriptions
  Lack auxiliary labels (e.g., edge categories)

[1] Collective classification in network data
[2] Towards Better Dynamic Graph Learning: New Architecture and Unified Library
[3] Open graph benchmark: Datasets for machine learning on graphs

Limitation of Existing Datasets

Limitation 1: Lack raw text descriptions of both nodes and edges

→ Bringing challenges to investigating the benefits of text attribute modeling on dynamic graph applications.

Limitation 2: Lack reasonable time segmentation

→ Brings challenges to investigating the semantic and structure co-evolution for dynamic graphs.

Limitation 3: Lack auxiliary labels

→ Brings challenges to investigating more valueable tasks on dynamic graphs.

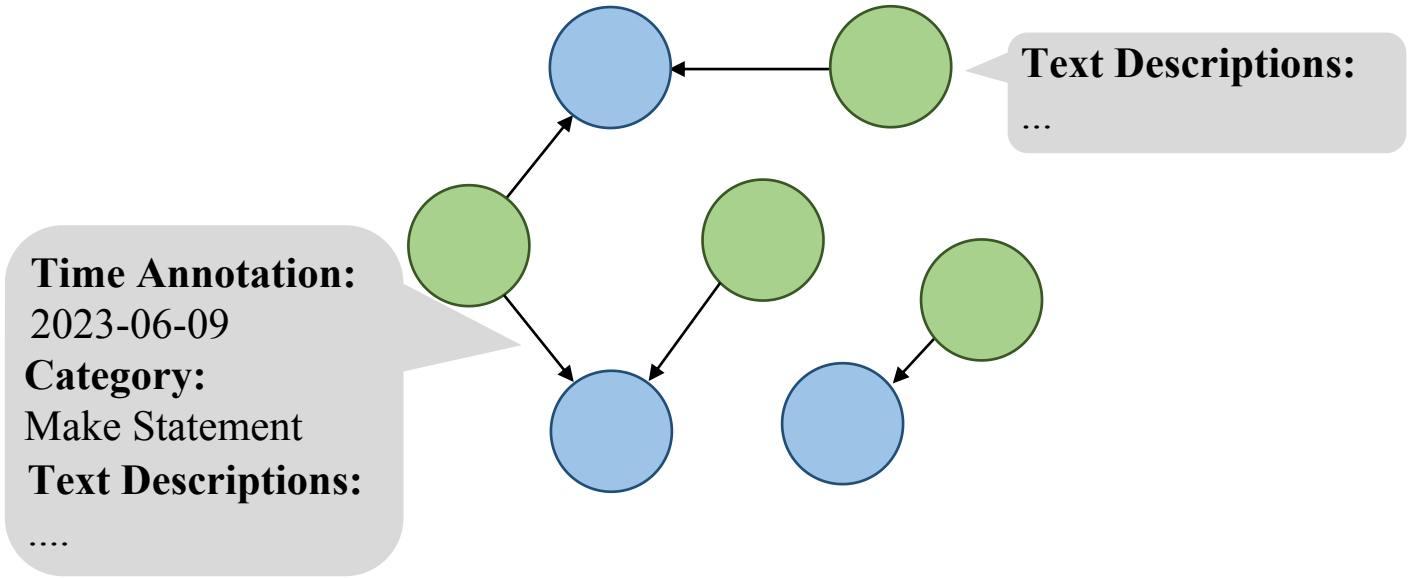
↓

Insufficent to describe complex interactions in real world

Fail to faithfully reflect the challenges in modeling real-world scenarios

The Proposed DTGB Datasets

DyTAG Formulation. A DyTAG can be defined as $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V}$ is the node set, $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ is the edge set. Let $\mathcal{T}$ denote the set of observed timestamps, $\mathcal{D}$, $\mathcal{R}$ and $\mathcal{L}$ are the set of node text descriptions, edge text descriptions, and edge categories, respectively. Each $v \in \mathcal{V}$ is associated with a text description $d_v \in \mathcal{D}$. Each $(u, v) \in \mathcal{E}$ can be represented as $(r_{u,v}, l_{u,v}, t_{u,v})$ with a text description $r_{u,v} \in \mathcal{R}$, a category $l_{u,v} \in \mathcal{L}$ and a timestamp $t_{u,v} \in \mathcal{T}$ to indicate the occurring time of this edge.



The Proposed DTGB Datasets

	Dataset	Nodes	Edges	Edge Categories	Timestamps	Domain	Text Attributes	Bipartite Graph
Previous Dynamic Graphs	**tgbn-trade**	255	468,245	N.A.	32	Trade	✗	✗
	tgbl-wiki	9,227	157,474	N.A.	152,757	Interaction	✗	✗
	tgbl-review	352,637	4,873,540	N.A.	6,865	E-commerce	✗	✓
	MOOC	7,144	411,749	N.A.	345,600	Interaction	✗	✓
	LastFM	1,980	1,293,103	N.A.	1,283,614	Interaction	✗	✓
Previous TAGs	**ogbn-arxiv-TA**	169,343	1,166,243	N.A.	N.A.	Academic	Node	✗
	CitationV8	1,106,759	6,120,897	N.A.	N.A.	Academic	Node	✗
	Books-Children	76,875	1,554,578	N.A.	N.A.	E-commerce	Node	✗
	Ele-Computers	87,229	721,081	N.A.	N.A.	E-commerce	Node	✗
	Sports-Fitness	173,055	1,773,500	N.A.	N.A.	E-commerce	Node	✗
Ours	**Enron**	42,711	797,907	10	1,006	E-mail	Node & Edge	✗
	GDELT	6,786	1,339,245	237	2,591	Knowledge graph	Node & Edge	✗
	ICEWS1819	31,796	1,100,071	266	730	Knowledge graph	Node & Edge	✗
	Stack elec	397,702	1,262,225	2	5,224	Multi-round dialogue	Node & Edge	✓
	Stack ubuntu	674,248	1,497,006	2	4,972	Multi-round dialogue	Node & Edge	✓
	Googlemap CT	111,168	1,380,623	5	55,521	E-commerce	Node & Edge	✓
	Amazon movies	293,566	3,217,324	5	7,287	E-commerce	Node & Edge	✓
	Yelp	2,138,242	6,990,189	5	6,036	E-commerce	Node & Edge	✓

https://github.com/zjs123/DTGB

First DyTAG Benchmark

- Diverse Domain
- Both Node&Edge Text
- Edge Lables
- Large Scale

Standardized Evaluation Protocol

- Future Link Prediction
- Edge Classification
- Destination Node Retrieval
- Textural Relation Generation
- Stadard Evaluation Pipeline

Empirical Observation

- Text is Helpful in Many Tasks
- Limited Edge Modeling
- Lack Fine-grained Structure-Semantic Evolution Modeling
- Scalability issue

Detailed Datasets Introduction

(1) E-commerce Datatsets (Amazon Movies; Googlemap_CT; Yelp)

Nodes: Users; Items

Edges: User review item

User Node Text: Name

Item Node Text: Title; Descrition; category

Edge Text: Reviews

Edge Labels: Ratings (1-5)

Item Node Text:
Title:"Praise Aerobics VHS"
Category: "Genre for Featured Categories", "Exercise & Fitness";
Description"Praise Aerobics - A low-intensity/high-intesity low impact aerobic workout." "Praise Aerobics VHS"

Edge Text: *So sorry I didn't purchase this years ago when it first came out!! This is very good and entertaining! We absolutely loved it and anticipate seeing it repeatedly. We actually wore out the cassette years back, so we also purchased this same product on cd. Best purchase we made out of all! Would purchase on dvd if we could find one.*

Edge Label: *5 (Very Good)*

Detailed Datasets Introduction

(2) Multi-round Dialogue Datatsets (Stack_ubuntu; Stack_elec)

Nodes: Users; Posts

Edges: User discuss on the posts

User Node Text: Name, self-introduce

Item Node Text: Title; Descrition

Edge Text: Answers or discussion from users

Edge Labels: Useless (0); Useful (1)

Post Node Text: *In my opinion this is one of those stuffy rules touted by grammarians who probably should have better things to do... When you can avoid it, don't end sentences with prepositions, but if rewriting the sentence will make it grammatically tortured, it's best to break the rule for the sake of clarity.*

Edge Text: *I really do hope this site remains unspoiled by these grammarians you mention. So far, it's good to see many advocating the breaking of rules where it feels sensible and natural.*

Edge Label: *1 (Useful)*

Detailed Datasets Introduction

(3) Temporal Knowledge Graphs (ICEWS1819; GDELT)

Nodes: Political entities

Edges: One entity has behavior with another

Node Text: Descrition of the entity

Edge Text: Descrition of their behavior

Edge Labels: Behavior types

Node Text: *Rodrigo Duterte: Executive,Government,Executive Office. Country is Philippines*

Lawmaker (Russia): Sector is Legislative / Parliamentary,Government. Country is Russian Federation

Edge Text: *Acknowledge or claim responsibility*

Edge Label: *MAKE PUBLIC STATEMENT*

Detailed Datasets Introduction

(4) E-mail Graph (Enron)

Nodes: Users

Edges: One user e-mail to another

Node Text: User name

Edge Text: Content of E-mail

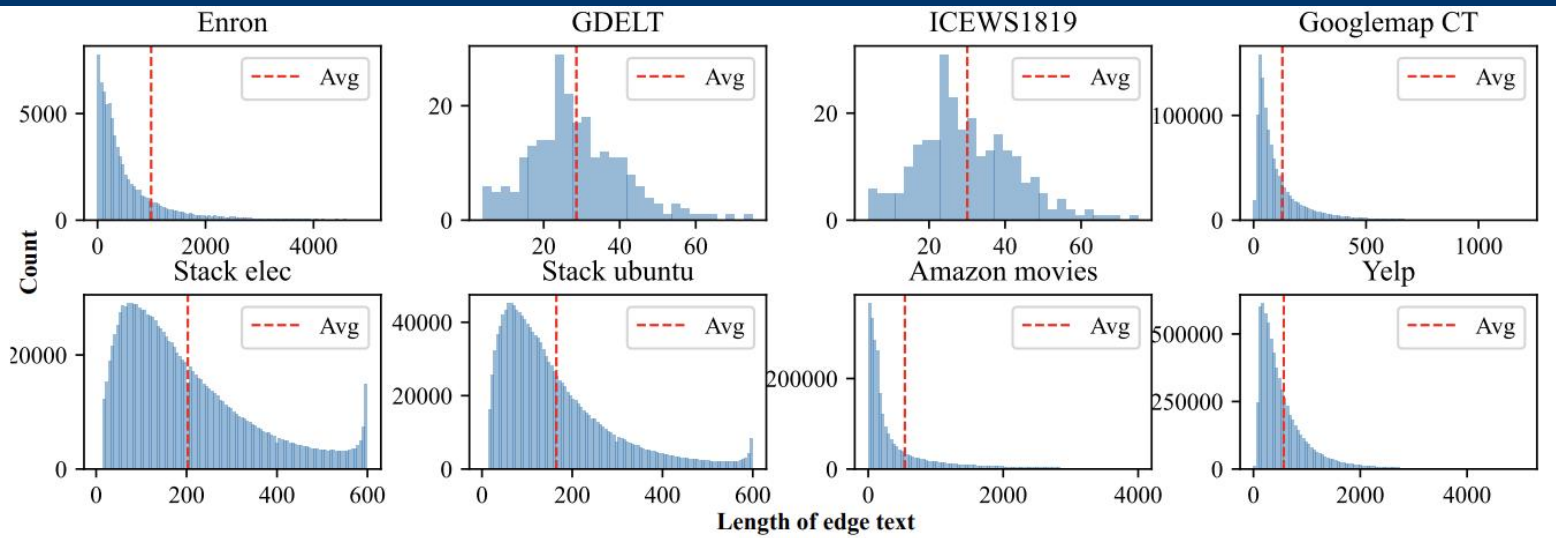
Edge Labels: E-mail archiving

Node Text:
marss@perkinscoie.com
dbkinnard@pplmt.com
brenda.herod@enron.com

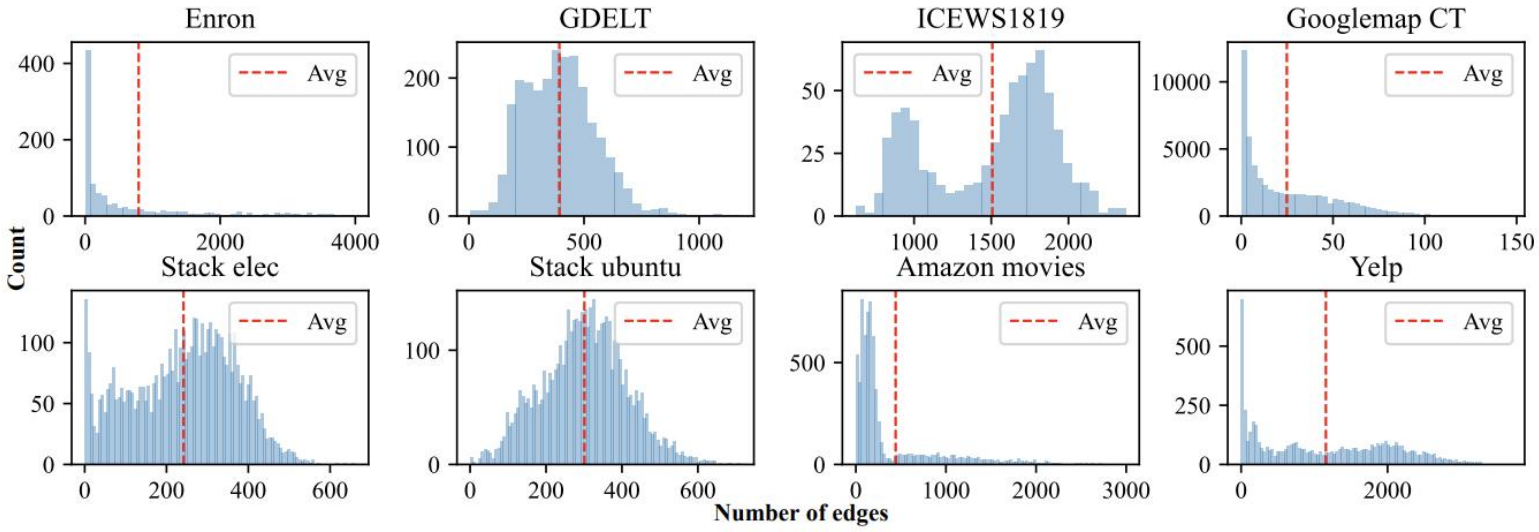
Edge Text: *Attached, please find the chaptered version SBX 43 relating to the San Diego rate freeze. I have also attached ABX 43 which is the companion measure to SBX 43. ABX 43 has not been signed by the Governor, however it is expected that he will sign the bill.*

Edge Label: *discussion_threads*

Datasets Statistics

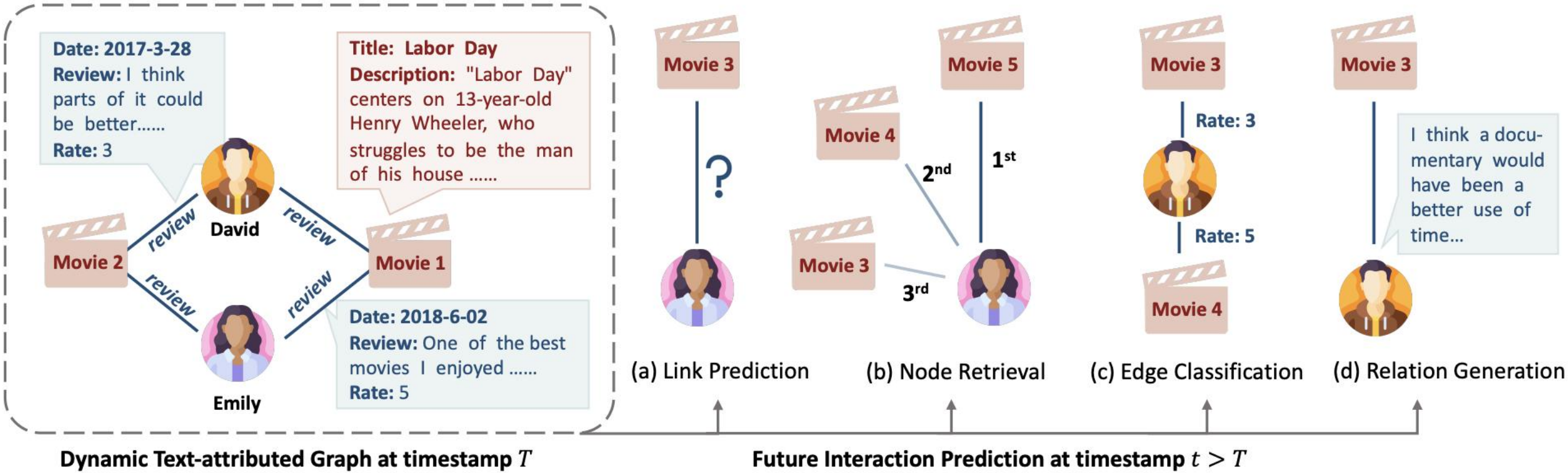


Distribution of edge text length



Distribution of the number of edges in each timestamp

Evaluation Protocol



Future Link Prediction

Table 3: AUC-ROC for future link prediction. *tr.* means transductive setting and *in.* means inductive setting. **Text** means whether to use Bert-encoded embeddings for initialization.

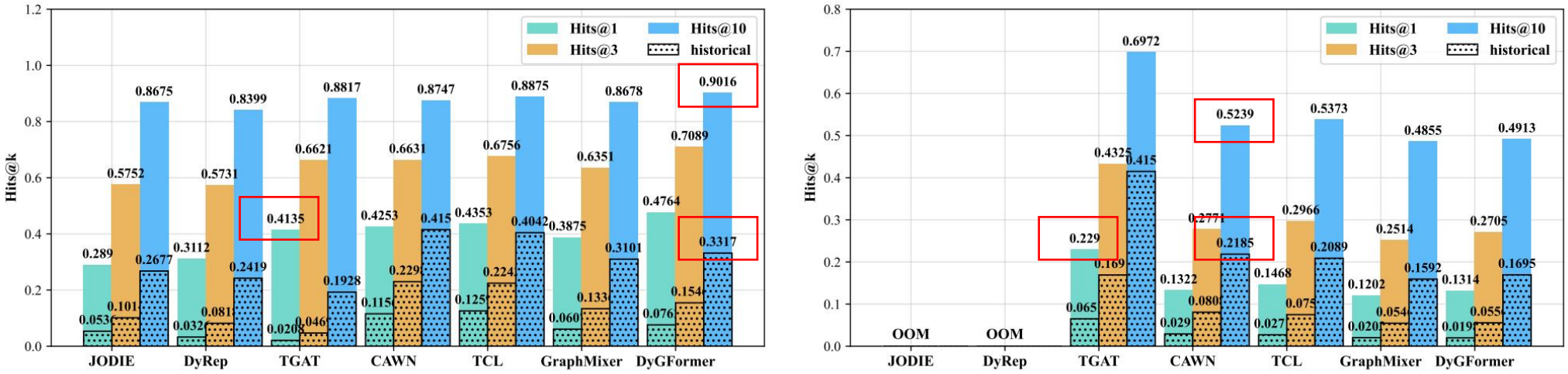
	Datasets	Text	JODIE	DyRep	TGAT	CAWN	TCL	GraphMixer	DyGFormer
tr.	Enron	✗	**0.9712 ± 0.0097**	0.9545 ± 0.0023	0.9511 ± 0.0011	0.9652 ± 0.0012	0.9604 ± 0.0079	0.9254 ± 0.0046	0.9653 ± 0.0015
		✓	0.9731 ± 0.0052	0.9274 ± 0.0026	0.9681 ± 0.0026	0.9740 ± 0.0007	0.9618 ± 0.0025	0.9567 ± 0.0013	**0.9779 ± 0.0014**
	ICEWS1819	✗	0.9821 ± 0.0095	0.9799 ± 0.0039	0.9787 ± 0.0065	0.9815 ± 0.0041	0.9842 ± 0.0036	0.9399 ± 0.0079	**0.9865 ± 0.0024**
		✓	0.9741 ± 0.0113	0.9632 ± 0.0027	0.9904 ± 0.0039	0.9857 ± 0.0018	**0.9923 ± 0.0012**	0.9863 ± 0.0024	0.9888 ± 0.0015
	Googlemap CT	✗	OOM	OOM	0.8537 ± 0.0153	**0.8543 ± 0.0027**	0.7740 ± 0.0013	0.7087 ± 0.0088	0.7864 ± 0.0047
		✓	OOM	OOM	**0.9049 ± 0.0071**	0.8687 ± 0.0063	0.8348 ± 0.0094	0.8095 ± 0.0014	0.8207 ± 0.0018
	GDELT	✗	0.9562 ± 0.0027	0.9477 ± 0.0011	0.9341 ± 0.0046	0.9419 ± 0.0026	0.9571 ± 0.0007	0.9316 ± 0.0021	**0.9648 ± 0.0007**
		✓	0.9533 ± 0.0020	0.9453 ± 0.0018	0.9595 ± 0.0033	0.9600 ± 0.0061	0.9619 ± 0.0008	0.9552 ± 0.0018	**0.9662 ± 0.0003**
in.	Enron	✗	0.8745 ± 0.0041	0.8560 ± 0.0124	0.8079 ± 0.0047	0.8710 ± 0.0030	0.8363 ± 0.0068	0.7510 ± 0.0071	**0.8991 ± 0.0012**
		✓	0.8732 ± 0.0037	0.7901 ± 0.0047	0.8650 ± 0.0032	0.9091 ± 0.0014	0.8512 ± 0.0062	0.8347 ± 0.0039	**0.9316 ± 0.0015**
	ICEWS1819	✗	0.9115 ± 0.0081	0.9390 ± 0.0054	0.9151 ± 0.0061	0.9330 ± 0.0076	0.9471 ± 0.0011	0.8858 ± 0.0089	**0.9613 ± 0.0010**
		✓	0.9285 ± 0.0065	0.9030 ± 0.0097	0.9706 ± 0.0054	0.9774 ± 0.0039	**0.9778 ± 0.0012**	0.9605 ± 0.0025	0.9630 ± 0.0027
	Googlemap CT	✗	OOM	OOM	0.7958 ± 0.0012	**0.7968 ± 0.0007**	0.7104 ± 0.0015	0.6675 ± 0.0033	0.7148 ± 0.0024
		✓	OOM	OOM	**0.8791 ± 0.0028**	0.7058 ± 0.0047	0.7895 ± 0.0046	0.7543 ± 0.0018	0.7648 ± 0.0052
	GDELT	✗	0.8977 ± 0.0035	0.8791 ± 0.0002	0.7501 ± 0.0074	0.7909 ± 0.0010	0.8544± 0.0045	0.7361 ± 0.0058	**0.9135± 0.0024**
		✓	0.8921 ± 0.0065	0.8917 ± 0.0007	0.9012 ± 0.0011	0.8899 ± 0.0082	0.9099 ± 0.0022	0.8942 ± 0.0035	**0.9206 ± 0.0003**

Simply using pre-trained embeddings to integrate the text information can result in performance degradation for memory-based models (e.g., DyRep)

Text information is helpful, especially in the inductive setting where test nodes are unseen during training.

Memory-based mrthods suffer from high consumption for large dynamic graphs.

Destination Node Retrieval



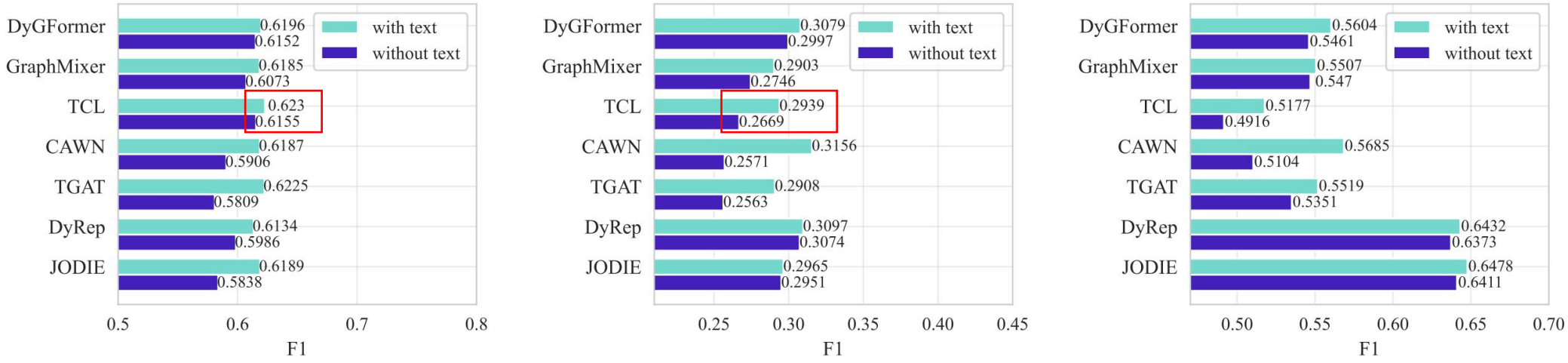
(a) GDELT (b) Googlemap CT

Figure 5: Node retrieval performance using random sampling and historical sampling.

Although existing models can achieve high accuracy on link prediction (more than 0.95), they still fail to get satisfactory perfromance on node retrieval

Existing models perform significantly worse in the historical sampling setting, showcasing these models largely rely on capturing structural and temporal co-occurrences, but ignore the semantic relevance.

Edge Classification



(a) Googlemap CT (b) ICEWS1819 (c) Enron

Figure 4: Edge classification performance with and without text attributes.

Existing models fail to achieve satisfactory performance on this task, especially on datasets with a large number of categories, because of their neglecting of edge information modeling

Text information consistently helps models achieve better performance on each datasets, verifying the necessity of integrating text attributes into temporal graph modeling

Textural Relation Generation

Description of item A: Name:xxx, Introduction: Family-run pizzeria with standout baby clam pies in a cozy space decorated with family photos.

Recent reviews of item A from other users:
1. My favorite pizza of all time!
2. Totally Great Pizza get the special.

Recent reviews of user P to other items:
1. item: Name: yyy, Introduction: ...
   review: The BBQ is fabulous. I has the #1 Combo with pulled pork, brisket, and ribs. AWESOME!
2. item: Name: zzz, Introduction: ...
   review: Excellent wine and spirits store.

Human Question: If User P visit Item A in the next time, please give me three possible review of User P to Item A.

LLM Response:
1. Delicious Pizza!!!
2. One of the few authentic apizza spots.
3. The sausage is their best pie, simply because it's made by them.

Figure 9: Example of prompts used for inference and fine-tuning in textual relation generation task (A case in `Googlemap CT` dataset).

Textural Relation Generation

Table 5: Precision, Recall and F1 of BERT Score of different LLMs on textural generation tasks. The number of test samples is 500 per dataset.

	Googlemap CT			Amazon Movies			Stack elec		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
GPT 3.5 turbo	79.89	**84.13**	81.91	79.79	83.61	81.63	80.52	81.96	81.21
GPT 4o	78.33	84.06	81.07	78.68	**84.20**	81.33	78.30	82.37	80.26
Llama3-8b	78.62	83.84	81.12	78.48	83.97	81.09	79.91	82.35	81.09
Mistral-7b	**80.21**	84.05	**82.07**	79.81	84.05	**81.84**	80.25	**82.61**	81.40
Vicuna-7b	80.04	83.79	81.85	**80.23**	83.60	81.83	**80.65**	82.37	**81.46**
Vicuna-13b	80.14	84.00	81.99	77.59	83.56	80.39	80.57	82.20	81.33

Table 6: Performance of LLMs after SFT on relation generation task in terms of BERT Score (F1)

	Googlemap CT	Stack elec
Llama3-8b	81.12	81.09
Llama3-8b + SFT	81.84 (0.72↑)	81.97 (0.88↑)
Vicuna-7b	81.85	81.46
Vicuna-7b + SFT	**85.67** (3.82↑)	82.67 (1.21↑)
Vicuna-13b	81.99	81.33
Vicuna-13b + SFT	84.67 (2.68↑)	**82.73** (1.40↑)

Open-source LLMs such as Mistral and Vicuna perform comparably well to proprietary LLMs in this task.

Supervised fine-tuning helps LLM to get better performance on this task.

Future Directions

(1) Controlled textual relation generation

Generate text between nodes with specific purpose (e.g., generating positive review from a user to a item).

(2) LLM for dynamic text-attributed graph reasoning

Temporal graph tokens that can directly incorporate long-range dynamic graph information into LLMs for reasoning and dynamics-aware generation

(3) Scalable and powerful dynamic text-attributed graph representation learning

Given the long text descriptions associated with nodes and edges, as well as long historical structure, how to efficently modeling their semantic and structure co-evolution

(4) Temporal question answering and evolution summarization on DyTAGs

Given the E-mail or discussion history among users, summary the development of thier conversation and answer questions about thier discussion via LLMs.