
BENCHTEMP: A General Benchmark for Evaluating Temporal Graph Neural Networks

Authors' Response to Reviewer hWRt

Opportunities For Improvement:

W1. original ideas on how to improve the evaluation setup for TGNN in general is needed. This work summarizes and standardizes existing evaluation settings however numerical results still show "controversial and inconsistent results" (using the sentence from the paper on page 2). For example on many datasets, SOTA methods can achieve > 95% AUROC and / or AP thus the ranking of the method still remains problematic.

W2. most if not all datasets are presented from prior literature in temporal graph learning. Given that this is a dataset and benchmark track submission, it would be helpful if the authors can contribute novel datasets in the benchmarking pipeline.

W3. Node reindexing is not a novel contribution and should only be an implementation detail. It is already utilized in prior work's implementation such as in the TGN source code, see the reindex function here.

W4. training time per epoch and # of epoch until early stopping are not good metrics for measuring efficiency. In many real world applications, the inference time of methods might be more important as they are deployed in the real world. In addition, the time per epoch is not meaningful unless the same number of epochs are measured and the # of epoch could be dependent on model hyperparameter, model initialization and parameter for early stopping thus not uniform. A simpler approach could just measure overall training time.

W5. Minor suggestions

1) Table 2 dataset statistics formatting looks a bit off due to the added equations, these can be explained in text. Taobao should be moved up with the datasets that are heterogenous if the table is to be ordered consistently

2) The dataset is hosted via google drive which is not a permanent storage option if the account was deactivated or lost, the datasets can no longer be accessed. I suggest hosting them on platform such as zenodo (<https://zenodo.org/>)

3) not sure what you mean by "We evaluate Reddit, Wikipedia, and MOOC datasets since they have two classes of node labels". You picked these datasets because the labels on them are two classes? Why not multi-class classification?

1

General Response:

2
3 Thanks for the valuable suggestion! Indeed, similar to the experimental results in prior literature,
4 on many datasets (**especially small datasets**), SOTA methods can achieve > 95% AUC-ROC or AP.
5 Thus, we have included new datasets with up to several million edges and nodes. We have added four
6 large-scale datasets (eBay-Large, Taobao-Large, DGraphFin, YouTubeReddit-Large).

7 The eBay datasets are a collection of the user transactions on eBay's e-commerce platform. We thank
8 our industrial collaborator for sharing their datasets in our research. Considering user privacy and
9 security, eBay datasets could only be shared among collaborators. Any researchers who are interested
10 in the eBay datasets, please email our team (jonnyhuanghnu@gmail.com). All datasets have been
11 hosted on the open-source platform zenodo(<https://zenodo.org/>) with a Digital Object Identifier (DOI)
12 10.5281/zenodo.8267771 (<https://zenodo.org/record/8267846>)).

Submitted to the 37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks. Do not distribute.

The experimental results on large-scale datasets may be more convincing. Furthermore, we have added Average Rank metric for ranking model performances on the newly added large-scale datasets for evaluating TGNN models.

We have added **the inference time** metric to evaluate the efficiency of methods. We have updated Table 2 of the paper (<https://openreview.net/pdf?id=rnZm2vQq31>).

We provide our response to each individual comment below:

Comment 1

W1. original ideas on how to improve the evaluation setup for TGNN in general is needed. This work summarizes and standardizes existing evaluation settings however numerical results still show "controversial and inconsistent results" (using the sentence from the paper on page 2). For example on many datasets, SOTA methods can achieve > 95% AUROC and / or AP thus the ranking of the method still remains problematic.

Response:

We thank the reviewer for the suggestions! In this paper, we present BenchTeMP, which unifies the pipeline of evaluating TGNN. We extensively compare TGNN models on dynamic link prediction and dynamic node classification tasks with diverse settings (transductive, inductive, inductive New-Old, and inductive New-New) and metrics (**runtime, memory, and inference time**).

Indeed, similar to the experimental results in prior literature [1–4], on many datasets, especially small datasets, SOTA methods can achieve > 95% AUC-ROC or AP. Thus, we have included new datasets with up to several million edges and nodes. We have added *six* datasets (eBay-Small, eBay-Large, Taobao-Large, DGraphFin, YouTubeReddit-Small, YouTubeReddit-Large), including *four large-scale* datasets (eBay-Large, Taobao-Large, DGraphFin, YouTubeReddit-Large). The statistics of the new datasets are shown in Table 1. The eBay datasets are a collection of the user transactions on eBay’s e-commerce platform. We thank our industrial collaborator for sharing their datasets in our research. Considering user privacy and security, eBay datasets could only be shared among collaborators. Any researchers who are interested in the eBay datasets, please email our team (jonnyhuanghnu@gmail.com). For easy access, all datasets have been hosted on the open-source platform zenodo with a Digital Object Identifier (DOI) 10.5281/zenodo.8267846 (<https://zenodo.org/record/8267846>).

The experimental results on large-scale datasets may be more convincing. Furthermore, we have added **Average Rank** metric for ranking model performances on the newly added large-scale datasets (eBay-Large, Taobao-Large, DGraphFin, YouTubeReddit-Large) to evaluate TGNN models on dynamic link prediction task and node classification task shown in Table 2 and Table 4.

- **eBay-Small** is a subset of the eBay-Large dataset. We sample 38,427 nodes and 384,677 edges from eBay-Large graph according to edge timestamps.
- **YouTubeReddit-Small** is a collection of massive visual contents on YouTube and long-term community activity on Reddit. This dataset covers a 3-month period from January to March 2020.

Table 1: Dataset statistics of the new datasets.

	<i>Domain</i>	<i># Nodes</i>	<i># Edges</i>
eBay-Small	E-commerce	38,427	384,677
YouTubeReddit-Small [5]	Social	264,443	297,732
eBay-Large	E-commerce	1,333,594	1,119,454
DGraphFin [6]	E-commerce	3,700,550	4,300,999
Youtube-Reddit-Large [5]	Social	5,724,111	4,228,523
Taobao-Large [2, 7]	E-commerce	1,630,453	5,008,745

- Each row in the dataset represents a YouTube video v_i being shared in a subreddit s_j by some user u_k at time t [5]. Nodes are YouTube videos and subreddits, edges are the users' interactions between videos and subreddits. This dynamic graph has 264,443 nodes and 297,732 edges.
- **eBay-Large** is a million-scale dataset consisting of 1.3 million nodes and 1.1 million edges, which comprises the selected transaction records from the eBay e-commerce platform over a two-month period. eBay-Large is modeled as a user-item graph, where items are heterogeneous entities which include information such as phone numbers, addresses, and email addresses associated with a transaction. We select one month of transactions as seed nodes and then expand each seed node two hops back in time to enrich the topology while maintaining consistency in the distribution of seed nodes.
 - **DGraphFin** is a collection of large-scale dynamic graph datasets, consisting of interactive objects, events and labels that evolve with time. It is a directed, unweighted dynamic graph consisting of millions of nodes and edges, representing a realistic user-to-user social network in financial industry. Nodes are users, and an edge from one user to another means that the user regards the other user as the emergency contact person [6].
 - **Youtube-Reddit-Large** dataset covers 54 months of YouTube video propagation history from January 2018 to June 2022 [5]. This dataset has 5,724,111 nodes and 4,228,523 edges.
 - **Taobao-Large** is a collection of the Taobao user behavior dataset intercepted based on the period 8:00 to 18:00 on 26 November 2017 [7]. Nodes are users and items, and edges are behaviors between users and items, such as favor, click, purchase, and add an item to shopping cart. This public dataset has 1,630,453 nodes and 5,008,74 user-item interaction edges.

A Experiments

We conduct extensive experiments on the tasks of *dynamic link prediction* and *dynamic node classification*. The experimental setup is the same as in the paper <https://openreview.net/pdf?id=rnZm2vQq31>.

A.1 Link Prediction Task

We run the link prediction task on 7 TGNN models and the new datasets under different settings (Transductive, Inductive, Inductive New-Old, and Inductive New-New). The AUC and AP results for each new datasets are shown in Table 2 and Table 3, respectively. For the four large-scale datasets (eBay-Large, Taobao-Large, DGraphFin, YouTubeReddit-Large), we observe the similar results as in the paper. Specifically, NAT and NeurTW achieve the top-2 performance on almost all datasets under transductive and inductive settings.

A.2 Node Classification Task

The eBay-Small and eBay-Large datasets have node labels, so we conduct dynamic node classification experiments on both the eBay-Small and eBay-Large datasets. The AUC results are shown in Table 4. We can observe the similar results as in the paper. NeurTW achieves the best performance on both eBay-Small and eBay-Large datasets. NAT performs poorly on the node classification task.

A.3 Efficiency

Considering many real world applications and , we add **the inference time** metric to evaluate the efficiency of models. The inference time comparison per 100,000 edges is shown in Figure 1. According to the figure, we can observe the similar model efficiency results as in the paper. In terms of the inference time, JODIE, DyRep, TGN and TGAT are faster, while CAWN and NeurTW are much slower. NAT is relatively faster than temporal walk-based methods through caching and parallelism optimizations, *achieving a good trade-off between model quality and efficiency*.

Table 2: ROC AUC results of new datasets on the *dynamic link prediction task*. The best and second-best results are highlighted as **bold red** and underlined blue. **Average Rank** are computed by the experimental results of models on four large-scale datasets (eBay-Large, Taobao-Large, DGraphFin, YouTubeReddit-Large). We do not highlight the second-best if the gap is > 0.05 compared with the best result.

Model \ Dataset	Transductive						
	JODIE	DyRep	TGN	TGAT	CAWN	NeurTW	NAT
eBay-Small	0.9946 \pm 0.0002	0.9941 \pm 0.0006	0.9984 \pm 0.0003	0.9838 \pm 0.0006	0.9985 \pm 0.0	0.9991 \pm 0.0	<u>0.9978 \pm 0.0003</u>
YouTubeReddit-Small	<u>0.8519 \pm 0.0007</u>	0.8499 \pm 0.0012	0.8432 \pm 0.0032	0.8441 \pm 0.0014	0.7586 \pm 0.0031	0.9003 \pm 0.0031	0.8259 \pm 0.005
eBay-Large	0.9614 \pm 0.0	0.9619 \pm 0.0001	<u>0.9642 \pm 0.0003</u>	0.5311 \pm 0.0003	0.9442 \pm 0.0003	0.9608 \pm 0.0	0.9658 \pm 0.0002
DGraphFin	0.8165 \pm 0.0024	0.8171 \pm 0.0016	0.8683 \pm 0.0023	0.6112 \pm 0.0165	0.5466 \pm 0.0103	<u>0.8611 \pm 0.0035</u>	0.8258 \pm 0.0001
Youtube-Reddit-Large	0.8532 \pm 0.0003	0.8529 \pm 0.0006	0.8458 \pm 0.0025	0.8536 \pm 0.0026	0.7466 \pm 0.0012	0.916 \pm 0.0025	<u>0.8605 \pm 0.0009</u>
Taobao-Large	0.7726 \pm 0.0005	0.7724 \pm 0.001	<u>0.8464 \pm 0.0008</u>	0.5567 \pm 0.0047	0.7771 \pm 0.0068	0.859 \pm 0.0091	0.8188 \pm 0.001
Average Rank	4.5	4.5	2.75	5.75	6	2.25	2.25
Model \ Dataset	Inductive						
	JODIE	DyRep	TGN	TGAT	CAWN	NeurTW	NAT
eBay-Small	0.9696 \pm 0.0007	0.9674 \pm 0.0018	0.9913 \pm 0.0004	0.9698 \pm 0.0006	0.9964 \pm 0.0001	<u>0.9982 \pm 0.0</u>	0.9998 \pm 0.0001
YouTubeReddit-Small	0.7582 \pm 0.0003	0.7545 \pm 0.0009	0.7276 \pm 0.0033	0.7436 \pm 0.0006	0.7533 \pm 0.0016	0.8978 \pm 0.0032	0.9876 \pm 0.0049
eBay-Large	0.7536 \pm 0.0014	0.7515 \pm 0.0006	0.7657 \pm 0.0026	0.5224 \pm 0.0003	0.9459 \pm 0.0001	<u>0.9608 \pm 0.0</u>	0.9999 \pm 0.0001
DGraphFin	0.6884 \pm 0.0051	0.6876 \pm 0.001	0.6439 \pm 0.0089	0.5677 \pm 0.0184	0.5479 \pm 0.009	0.8635 \pm 0.0021	<u>0.7955 \pm 0.0201</u>
Youtube-Reddit-Large	0.7539 \pm 0.0005	0.7554 \pm 0.0003	0.7243 \pm 0.0016	0.7501 \pm 0.0019	0.7327 \pm 0.0016	<u>0.9128 \pm 0.0031</u>	0.9863 \pm 0.006
Taobao-Large	0.7075 \pm 0.0009	0.7042 \pm 0.0006	0.6812 \pm 0.0032	0.5222 \pm 0.0041	0.7787 \pm 0.0103	<u>0.869 \pm 0.010</u>	0.9933 \pm 0.0008
Average Rank	4	4.5	5.5	6.25	4.75	1.75	1.25
Model \ Dataset	Inductive New-Old						
	JODIE	DyRep	TGN	TGAT	CAWN	NeurTW	NAT
eBay-Small	0.9862 \pm 0.0003	0.9836 \pm 0.0016	0.9947 \pm 0.0009	0.9712 \pm 0.002	0.9985 \pm 0.0	<u>0.9988 \pm 0.0</u>	0.9999 \pm 0.0
YouTubeReddit-Small	0.7695 \pm 0.001	0.7655 \pm 0.0018	0.7396 \pm 0.0034	0.7242 \pm 0.0004	0.7573 \pm 0.0022	<u>0.922 \pm 0.0002</u>	0.9967 \pm 0.0014
eBay-Large	0.6109 \pm 0.0244	0.5906 \pm 0.0087	0.8134 \pm 0.0105	0.6363 \pm 0.0605	<u>0.9569 \pm 0.0007</u>	0.8973 \pm 0.0	1.0 \pm 0.0
DGraphFin	0.5768 \pm 0.0071	0.5735 \pm 0.0007	0.5564 \pm 0.0021	0.5742 \pm 0.013	0.5646 \pm 0.0244	<u>0.7702 \pm 0.0043</u>	0.8693 \pm 0.0066
Youtube-Reddit-Large	0.7844 \pm 0.0015	0.7894 \pm 0.0017	0.7623 \pm 0.0031	0.7457 \pm 0.0062	0.7511 \pm 0.0022	<u>0.9356 \pm 0.0004</u>	0.9958 \pm 0.0025
Taobao-Large	0.7023 \pm 0.0015	0.6953 \pm 0.0022	0.6771 \pm 0.0055	0.5104 \pm 0.0106	0.7674 \pm 0.005	<u>0.8458 \pm 0.0043</u>	0.9965 \pm 0.0005
Average Rank	4.25	5	5.5	5.75	4.25	2.25	1
Model \ Dataset	Inductive New-New						
	JODIE	DyRep	TGN	TGAT	CAWN	NeurTW	NAT
eBay-Small	0.9388 \pm 0.0009	0.9366 \pm 0.0037	0.9838 \pm 0.0007	0.9556 \pm 0.0007	0.9937 \pm 0.0	<u>0.9975 \pm 0.0</u>	0.9997 \pm 0.0004
YouTubeReddit-Small	0.7436 \pm 0.0015	0.7436 \pm 0.0018	0.7265 \pm 0.0055	0.749 \pm 0.0011	0.7479 \pm 0.004	<u>0.864 \pm 0.0071</u>	0.9868 \pm 0.0049
eBay-Large	0.7526 \pm 0.0013	0.7500 \pm 0.0005	0.7639 \pm 0.0027	0.5196 \pm 0.0002	0.9542 \pm 0.0003	<u>0.9615 \pm 0.0</u>	0.9999 \pm 0.0001
DGraphFin	0.7307 \pm 0.0007	0.7323 \pm 0.0002	0.6843 \pm 0.0131	0.5649 \pm 0.0248	0.5417 \pm 0.0099	0.9051 \pm 0.0028	<u>0.7584 \pm 0.0323</u>
Youtube-Reddit-Large	0.6932 \pm 0.0026	0.7022 \pm 0.0007	0.6703 \pm 0.0024	0.7269 \pm 0.0	0.6942 \pm 0.0028	<u>0.8716 \pm 0.0077</u>	0.9796 \pm 0.0103
Taobao-Large	0.7243 \pm 0.0001	0.7247 \pm 0.0001	0.6885 \pm 0.0024	0.5256 \pm 0.0054	0.7922 \pm 0.0118	<u>0.8906 \pm 0.0088</u>	0.9969 \pm 0.0002
Average Rank	5	4.25	5.5	5.75	4.5	1.75	1.25
Model \ Dataset	Inductive New-New						
	JODIE	DyRep	TGN	TGAT	CAWN	NeurTW	NAT
Total Rank	4.44	4.56	4.81	5.88	4.88	<u>2.00</u>	1.44

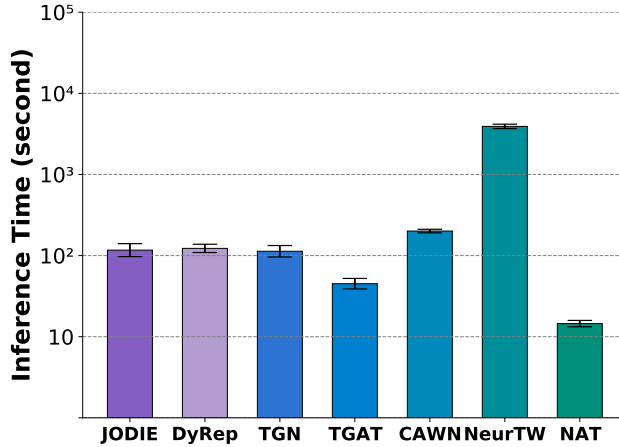


Figure 1: Inference time comparison per 100,000 edges.

Comment 2

W2. most if not all datasets are presented from prior literature in temporal graph learning. Given that this is a dataset and benchmark track submission, it would be helpful if the authors can contribute novel datasets in the benchmarking pipeline.

Table 3: AP results of new datasets on the *dynamic link prediction task*. The best and second-best results are highlighted as **bold red** and underlined blue. We do not highlight the second-best if the gap is > 0.05 compared with the best result.

		Transductive						
Model \ Dataset	JODIE	DyRep	TGN	TGAT	CAWN	NeurTW	NAT	
eBay-Small	0.9938 ± 0.0004	0.9936 ± 0.0006	<u>0.9983 ± 0.0003</u>	0.9819 ± 0.0009	0.9981 ± 0.0	0.9991 ± 0.0	0.9975 ± 0.0002	
YouTubeReddit-Small	<u>0.8612 ± 0.0009</u>	0.8594 ± 0.0012	0.8421 ± 0.0041	0.8515 ± 0.0012	0.7625 ± 0.0042	0.9112 ± 0.0021	0.8325 ± 0.0068	
eBay-Large	0.9318 ± 0.0002	0.9322 ± 0.0002	<u>0.9357 ± 0.0006</u>	0.5239 ± 0.0002	0.9144 ± 0.0004	0.9307 ± 0.0	0.9398 ± 0.0004	
DGraphFin	0.7705 ± 0.0009	0.7705 ± 0.0024	<u>0.8571 ± 0.0009</u>	0.6441 ± 0.0123	0.5431 ± 0.0095	0.8637 ± 0.0014	0.7956 ± 0.0012	
Youtube-Reddit-Large	0.8622 ± 0.0007	<u>0.8632 ± 0.0004</u>	0.8476 ± 0.0022	0.8591 ± 0.0026	0.7475 ± 0.0017	0.9222 ± 0.0013	0.8628 ± 0.0015	
Taobao-Large	0.7164 ± 0.0003	0.7142 ± 0.0008	<u>0.844 ± 0.0011</u>	0.5761 ± 0.0023	0.7616 ± 0.0069	0.8568 ± 0.016	0.7904 ± 0.0008	
		Inductive						
eBay-Small	0.9638 ± 0.0007	0.9619 ± 0.0017	0.9898 ± 0.0005	0.9675 ± 0.0007	0.9953 ± 0.0002	<u>0.9982 ± 0.0</u>	0.9998 ± 0.0001	
YouTubeReddit-Small	0.7866 ± 0.0007	0.7833 ± 0.0009	0.7387 ± 0.0069	0.7551 ± 0.0002	0.7568 ± 0.0031	<u>0.9086 ± 0.0022</u>	0.9872 ± 0.0056	
eBay-Large	0.6989 ± 0.0018	0.6973 ± 0.0007	0.7096 ± 0.0030	0.518 ± 0.0002	0.9174 ± 0.0001	<u>0.9308 ± 0.0</u>	0.9999 ± 0.0001	
DGraphFin	0.6563 ± 0.002	0.6567 ± 0.0009	0.624 ± 0.006	0.5866 ± 0.0123	0.5428 ± 0.0082	0.8626 ± 0.0012	<u>0.7053 ± 0.0185</u>	
Youtube-Reddit-Large	0.7796 ± 0.0009	0.7818 ± 0.0009	0.73 ± 0.0029	0.7587 ± 0.0025	0.7353 ± 0.0022	<u>0.9192 ± 0.0022</u>	0.9849 ± 0.0071	
Taobao-Large	0.6763 ± 0.0011	0.6746 ± 0.0011	0.6664 ± 0.0012	0.5315 ± 0.0027	0.7533 ± 0.011	<u>0.8596 ± 0.0205</u>	0.9941 ± 0.0007	
		Inductive New-Old						
eBay-Small	0.9849 ± 0.0007	0.9836 ± 0.0013	0.9931 ± 0.0008	0.9682 ± 0.0028	0.9985 ± 0.0001	<u>0.999 ± 0.0</u>	0.9999 ± 0.0	
YouTubeReddit-Small	0.7963 ± 0.0013	0.7937 ± 0.0014	0.729 ± 0.0086	0.7296 ± 0.0013	0.762 ± 0.0041	<u>0.9244 ± 0.0015</u>	0.9966 ± 0.0016	
eBay-Large	0.5670 ± 0.0186	0.5870 ± 0.0074	0.8024 ± 0.0060	0.6504 ± 0.0385	<u>0.9592 ± 0.0008</u>	0.8458 ± 0.0	1.0 ± 0.0	
DGraphFin	0.6005 ± 0.0048	0.5872 ± 0.0059	0.5753 ± 0.0062	0.5927 ± 0.0058	0.5669 ± 0.0269	<u>0.7572 ± 0.0025</u>	0.8184 ± 0.0088	
Youtube-Reddit-Large	0.808 ± 0.0014	0.8142 ± 0.0019	0.7472 ± 0.0043	0.7526 ± 0.0097	0.7553 ± 0.0025	<u>0.9368 ± 0.0009</u>	0.9953 ± 0.0028	
Taobao-Large	0.7009 ± 0.0013	0.698 ± 0.0014	0.6879 ± 0.0008	0.5254 ± 0.0074	0.7597 ± 0.0053	<u>0.8459 ± 0.0103</u>	0.9969 ± 0.0004	
		Inductive New-New						
eBay-Small	0.923 ± 0.001	0.9226 ± 0.0024	0.98 ± 0.0007	0.9505 ± 0.0009	0.991 ± 0.0001	<u>0.9973 ± 0.0</u>	0.9997 ± 0.0004	
YouTubeReddit-Small	0.7578 ± 0.0015	0.7582 ± 0.0021	0.7564 ± 0.0043	0.7718 ± 0.0023	0.7498 ± 0.004	<u>0.8868 ± 0.0034</u>	0.9861 ± 0.0063	
eBay-Large	0.6976 ± 0.0016	0.6957 ± 0.0007	0.7078 ± 0.0031	0.5154 ± 0.0001	0.93 ± 0.0003	<u>0.9318 ± 0.0</u>	0.9999 ± 0.0001	
DGraphFin	0.6802 ± 0.0005	0.6811 ± 0.0002	0.6526 ± 0.0098	0.5831 ± 0.0184	0.5379 ± 0.0071	0.8977 ± 0.0014	0.6529 ± 0.0249	
Youtube-Reddit-Large	0.7038 ± 0.0024	0.7115 ± 0.0007	0.6979 ± 0.002	0.7414 ± 0.0012	0.6965 ± 0.004	<u>0.8848 ± 0.0023</u>	0.9761 ± 0.0134	
Taobao-Large	0.6738 ± 0.0005	0.6742 ± 0.0005	0.6611 ± 0.0011	0.53 ± 0.0023	0.7521 ± 0.0127	<u>0.8738 ± 0.0145</u>	0.9973 ± 0.0001	

Table 4: ROC AUC results for the *dynamic node classification task* on the eBay datasets. The top-2 results are highlighted as **bold red** and underlined blue.

Model \ Dataset	JODIE	DyRep	TGN	TGAT	CAWN	NeurTW	NAT
eBay-Small	0.9274 \pm 0.0017	0.8677 \pm 0.0356	0.913 \pm 0.0025	<u>0.9342 \pm 0.0002</u>	0.9305 \pm 0.0001	0.9529 \pm 0.0002	0.6797 \pm 0.0115
eBay-Large	0.7244 \pm 0.0002	0.7246 \pm 0.0	0.6586 \pm 0.0129	0.672 \pm 0.0016	<u>0.7710 \pm 0.0002</u>	0.7859 \pm 0.0	0.5304 \pm 0.0011
Average Rank	4	4.5	5.5	3.5	<u>2.5</u>	1	7

90 **Response:**

91 Thanks for the valuable suggestion! We have included new datasets with up to several million edges
 92 and nodes. We have carefully thought through your comments and added six datasets (eBay-Small,
 93 eBay-Large, Taobal-Large, DGraphFin, YouTubeReddit-Small, YouTubeReddit-Large), including
 94 four large-scale datasets (eBay-Large, Taobao-Large, DGraphFin, YouTubeReddit-Large) shown in
 95 Table 1. The eBay datasets are a collection of the user transactions on eBay’s e-commerce platform.
 96 We thank our industrial collaborator for sharing their datasets in our research. Considering user
 97 privacy and security, eBay datasets could only be shared among collaborators. Any researchers
 98 who are interested in the eBay datasets, please email our team. All datasets have been hosted
 99 on the open-source platform zenodo(<https://zenodo.org/>) with a Digital Object Identifier (DOI)
 100 10.5281/zenodo.8267771 (<https://zenodo.org/record/8267846>)).

101 We have reported the corresponding experiments and detailed discussions in Section A.

Comment 3

W3. Node reindexing is not a novel contribution and should only be an implementation detail. It is already utilized in prior work’s implementation such as in the TGN source code, see the reindex function here.

102

103 **Response:**

104 We appreciate this suggestion! Indeed, node reindexing is an necessary implementation de-
105 tail. However, unlike previous codes in prior works' implementation, the node reindexing oper-
106 ation in BenchTeMP takes into account whether the graph is bipartite or not. Line 121-127 at
107 python file ([https://github.com/qianghuangwhu/benchtemp/blob/master/preprocess/](https://github.com/qianghuangwhu/benchtemp/blob/master/preprocess/preprocessing.py)
108 [preprocessing.py](https://github.com/qianghuangwhu/benchtemp/blob/master/preprocess/preprocessing.py)) of BenchTeMP shows the detail of implementation. Node reindexing is a
109 necessary operation for constructing temporal graph datasets. Thus, it is meaningful that BenchTeMP
110 improves the node reindexing operation and made it standard and add it into BenchTeMP PyPI library
111 (<https://pypi.org/project/benchtemp/>).

Comment 4

W4. training time per epoch and # of epoch until early stopping are not good metrics for measuring efficiency. In many real world applications, **the inference time** of methods might be more important as they are deployed in the real world. In addition, the time per epoch is not meaningful unless the same number of epochs are measured and the # of epoch could be dependent on model hyperparameter, model initialization and parameter for early stopping thus not uniform. A simpler approach could just measure overall training time.

112

113 **Response:**

114 Thanks for this valuable suggestion!
115 we have added **the inference time** metric to evaluate the efficiency of TGNN models. See Section
116 A.3 for details.

Comment 5

W5. Minor suggestions
Table 2 dataset statistics formatting looks a bit off due to the added equations, these can be explained in text. Taobao should be moved up with the datasets that are heterogenous if the table is to be ordered consistently

117

118 **Response:**

119 We appreciate the suggestion and totally agree! We have updated Table 2 of the paper (<https://openreview.net/pdf?id=rnZm2vQq31>).
120

Comment 6

W6. Minor suggestions
The dataset is hosted via google drive which is not a permanent storage option if the account was deactivated or lost, the datasets can no longer be accessed. I suggest hosting them on platform such as zenodo (<https://zenodo.org/>)

121

122 **Response:**

123 We appreciate your valuable suggestion! All datasets have been hosted on the open-source platform
124 **zenodo** with a Digital Object Identifier (DOI) **10.5281/zenodo.8267771** ([https://zenodo.org/](https://zenodo.org/record/8267846)
125 [record/8267846](https://zenodo.org/record/8267846)).

Comment 7

W7. Minor suggestions

not sure what you mean by "We evaluate Reddit, Wikipedia, and MOOC datasets since they have two classes of node labels". You picked these datasets because the labels on them are two classes? Why not multi-class classification?

126

127 **Response:**

128 We are grateful for this suggestion! Similar to the experiments of dynamic node classification in prior
129 literature, only Reddit, Wikipedia, and MOOC datasets have node labels (0 and 1). Datasets are intro-
130 duced at Section A of the Appendix ([https://openreview.net/attachment?id=rnZm2vQq31&](https://openreview.net/attachment?id=rnZm2vQq31&name=supplementary_material)
131 [name=supplementary_material](https://openreview.net/attachment?id=rnZm2vQq31&name=supplementary_material)). It is worth mentioning that the newly added eBay datasets
132 (eBay-Small, eBay-Large) have node labels and can be perform dynamic node classification task
133 shown in Section A.2. The experimental results for dynamic node classification on eBay datasets
134 Section A.2. The eBay datasets are a collection of the user transactions on eBay's e-commerce
135 platform. We thank our industrial collaborator for sharing their datasets in our research. Considering
136 user privacy and security, eBay datasets could only be shared among collaborators. Any researchers
137 who are interested in the eBay datasets, please email our team (jonnyhuanghnu@gmail.com).

References

- [1] Da Xu, Chuanwei Ruan, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. Inductive representation learning on temporal graphs. In *International Conference on Learning Representations*, 2020.
- [2] Ming Jin, Yuan-Fang Li, and Shirui Pan. Neural temporal walks: Motif-aware representation learning on continuous-time dynamic graphs. In *Advances in Neural Information Processing Systems*, 2022.
- [3] Yuhong Luo and Pan Li. Neighborhood-aware scalable temporal network representation learning. In *The First Learning on Graphs Conference*, 2022.
- [4] Farimah Poursafaei, Andy Huang, Kellin Pelrine, and Reihaneh Rabbany. Towards better evaluation for dynamic link prediction. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.
- [5] Yiqiao Jin, Yeon-Chang Lee, Kartik Sharma, Meng Ye, Karan Sikka, Ajay Divakaran, and Srijan Kumar. Predicting information pathways across online communities. *arXiv preprint arXiv:2306.02259*, 2023.
- [6] Xuanwen Huang, Yang Yang, Yang Wang, Chunping Wang, Zhisheng Zhang, Jiarong Xu, Lei Chen, and Michalis Vazirgiannis. Dgraph: A large-scale financial dataset for graph anomaly detection. *Advances in Neural Information Processing Systems*, 35:22765–22777, 2022.
- [7] Han Zhu, Xiang Li, Pengye Zhang, Guozheng Li, Jie He, Han Li, and Kun Gai. Learning tree-based deep model for recommender systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1079–1088, 2018.