BenchTemp: A General Benchmark for Evaluating Temporal Graph Neural Networks

Authors' Response to Reviewer hWRt - Second Round

Opportunities For Improvement:

One concern is that the eBay datasets (which is also the only datasets with node labels of the new ones added) is not shared publicly yet. The authors mentioned that it can only be shared with collaborators. I wonder if there is any plans to make the dataset public in the future. I understand that there might be delays or difficulties in sharing datasets, however as this is a part of the contribution, it would be important for the general public to have access to and benchmark on the eBay datasets.

For W1, novel datasets and the average rank are indeed interesting. However, it remains concerning that methods such as NAT and NeurTW already achieved 90%+ or even 95%+ AUC and AP performance on many of these datasets for both transductive and inductive settings thus leaving little room for improvement for future methods. Are the authors suggesting average rank should be the metric to rank these methods in terms of performance? There should also be more discussion regarding average rank in the main paper in this case. Regardless, I believe better evaluation setup and or metric is still needed. But might be left to future future. The authors have adequately addressed my other concerns.

General Response:

Dear Reviewer hWRt:

We sincerely appreciate your feedback and valuable comments!

We have carefully thought through your concerns about the open source of the eBay datasets and the over-optimistic evaluation on many of datasets.

After a discussion with our industrial partner eBay, we are working on sharing the **eBay-Small** and **eBay-Large** datasets in a way that ensures availability and justifies the research purpose:

1. We will build a website that describes the eBay datasets and provides an application form.
2. The applicants input their email and affiliation in the form, and agree to the access terms (similar to ImageNet).
3. The backend will check the applicant's information and send a download link to the corresponding email.

In the meantime, eBay provide a Google form for the applicants to obtain the eBay datasets: `https://forms.gle/bP1RmyVJ1C6pgyS66` (**the applicants can remain anonymous**).

Average Rank metric has been widely used in the SOTA benchmarks and should be the metric to rank these methods in terms of performance. we will add *Average Rank* into the main paper.

As for NAT and NeurTW can achieved 90%+ or even 95%+ AUC and AP performance on many of these datasets, we have found that BenchTeMP with a *Historical Negative Sampling* or *Inductive Negative Sampling* strategy can addressed this issue. We have conducted experiments of NAT on

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

those over-performance datasets with *Historical Negative Sampling* and *Inductive Negative Sampling*. The experimental results demonstrated the effectiveness of textitHistorical Negative Sampling and *Inductive Negative Sampling*.

Thank you and best regards!

Yours sincerely,

Qiang Huang, Jiawei Jiang, Xi Susie Rao, Ce Zhang, Zhichao Han, Zitao Zhang, Xin Wang, Quanqing Xu, Yang Zhao, Chuang Hu, Shuo Shang, Yongjun He, Bo Du

We provide our response to each individual comment below:

Comment 1

W1. One concern is that the eBay datasets (which is also the only datasets with node labels of the new ones added) is not shared publicly yet. The authors mentioned that it can only be shared with collaborators. I wonder if there is any plans to make the dataset public in the future. I understand that there might be delays or difficulties in sharing datasets, however as this is a part of the contribution, it would be important for the general public to have access to and benchmark on the eBay datasets.

Response:

We thank the reviewer for the suggestions! After a discussion with our industrial partner eBay, we are working on sharing the **eBay-Small** and **eBay-Large** datasets in a way that ensures availability and justifies the research purpose:

1. We will build a website that describes the eBay datasets and provides an application form.
2. The applicants input their email and affiliation in the form, and agree to the access terms (similar to ImageNet).
3. The backend will check the applicant's information and send a download link to the corresponding email.

Note that, many large-scale datasets also adopt this routine, e.g., YFCC100M from Yahoo (`http://www.multimediacommons.org/`) and ImageNet (`https://www.image-net.org/download.php`). We hope this solution can address the reviewer's concern. We will optimize this procedure according to the reviewer's further suggestions.

In the meantime, eBay provide a Google form for the applicants to obtain the eBay datasets: `https://forms.gle/bP1RmyVJ1C6pgyS66` (**the applicants can remain anonymous**).

Comment 2

W2. For W1, novel datasets and the average rank are indeed interesting. However, it remains concerning that methods such as NAT and NeurTW already achieved 90%+ or even 95%+ AUC and AP performance on many of these datasets for both transductive and inductive settings thus leaving little room for improvement for future methods. Are the authors suggesting average rank should be the metric to rank these methods in terms of performance? There should also be more discussion regarding average rank in the main paper in this case.

Response:

1. Average Rank

We appreciate your valuable suggestion! *Average Rank* metric has been widely used in the SOTA models and benchmarks and should be the metric to rank these methods in terms of performance.

We have computed the *Average Rank* metric of the Table 3 and Table 5 in the main paper (`https://openreview.net/pdf?id=rnZm2vQq31`), as shown in the tables below.

Table 1: ROC AUC results on the link prediction task. "*" denotes that the model encounters runtime error; "—" denotes timeout after 48 hours. The best and second-best results are highlighted as **bold red** and underlined blue. Some standard deviations are zero because we terminate those models that can only run one epoch within 2 days. 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
Reddit	0.9760 ± 0.0006	0.9803 ± 0.0005	0.9871 ± 0.0001	0.981 ± 0.0002	**0.9889 ± 0.0002**	0.9841 ± 0.0016	0.9854 ± 0.002
Wikipedia	0.9505 ± 0.0032	0.9426 ± 0.0007	0.9846 ± 0.0003	0.9509 ± 0.0017	0.9889 ± 0.0	**0.9912 ± 0.0001**	0.9786 ± 0.0035
MOOC	0.7899 ± 0.0208	0.8243 ± 0.0323	0.8999 ± 0.0213	0.7391 ± 0.0056	**0.9459 ± 0.0008**	0.8071 ± 0.0193	0.7568 ± 0.0305
LastFM	0.6766 ± 0.0590	0.6793 ± 0.0553	0.7743 ± 0.0256	0.5094 ± 0.0071	**0.8746 ± 0.0013**	0.839 ± 0.0	0.8536 ± 0.0027
Enron	0.8293 ± 0.0148	0.7986 ± 0.0358	0.8621 ± 0.0173	0.6161 ± 0.0214	0.9159 ± 0.0032	0.8956 ± 0.0045	**0.9212 ± 0.0029**
SocialEvo	0.8666 ± 0.0233	0.9020 ± 0.0026	**0.952 ± 0.0003**	0.7851 ± 0.0047	0.9337 ± 0.0003	—	0.9202 ± 0.0065
UCI	0.8786 ± 0.0017	0.5086 ± 0.0651	0.8875 ± 0.0161	0.7998 ± 0.0052	0.9189 ± 0.0017	**0.9670 ± 0.0031**	0.9076 ± 0.0116
CollegeMsg	0.5730 ± 0.0690	0.5382 ± 0.0058	0.8419 ± 0.084	0.8084 ± 0.0032	0.9156 ± 0.004	**0.9698 ± 0.0**	0.9059 ± 0.0122
CanParl	0.7939 ± 0.0063	0.7737 ± 0.0255	0.7575 ± 0.0694	0.7077 ± 0.0218	0.7197 ± 0.0905	**0.8920 ± 0.0173**	0.6917 ± 0.0722
Contact	0.9379 ± 0.0073	0.9276 ± 0.0206	0.9769 ± 0.0032	0.5582 ± 0.009	0.9685 ± 0.0028	**0.984 ± 0.0**	0.9463 ± 0.021
Flights	0.9449 ± 0.0073	0.8981 ± 0.0056	0.9787 ± 0.0025	0.9016 ± 0.0027	**0.9861 ± 0.0002**	0.9302 ± 0.0	0.9747 ± 0.0061
UNTrade	0.6786 ± 0.0103	0.6377 ± 0.0032	0.6543 ± 0.01	*	0.7511 ± 0.0012	0.5924 ± 0.0368	**0.783 ± 0.0472**
USLegis	0.8278 ± 0.0024	0.7425 ± 0.0374	0.8137 ± 0.002	0.7738 ± 0.0062	0.9643 ± 0.0043	**0.9715 ± 0.0009**	0.782 ± 0.0261
UNVote	0.6523 ± 0.0082	0.6236 ± 0.0305	**0.7176 ± 0.0109**	0.5134 ± 0.0026	0.6037 ± 0.0019	0.5871 ± 0.0	0.6776 ± 0.0411
Taobao	0.8405 ± 0.0006	0.8409 ± 0.0012	0.8654 ± 0.0005	0.5396 ± 0.009	0.7708 ± 0.0026	0.8759 ± 0.0009	**0.8937 ± 0.0015**
Average Rank	4.47	5.07	2.8	6	**2.47**	2.87	3
				Inductive			
Reddit	0.9514 ± 0.0045	0.9583 ± 0.0004	0.976 ± 0.0002	0.9651 ± 0.0002	0.9868 ± 0.0002	0.9802 ± 0.0011	**0.9906 ± 0.0034**
Wikipedia	0.9305 ± 0.0020	0.9099 ± 0.0031	0.9781 ± 0.0005	0.9343 ± 0.0031	0.989 ± 0.0003	0.9904 ± 0.0002	**0.9962 ± 0.0026**
MOOC	0.7779 ± 0.0575	0.8269 ± 0.0182	0.8869 ± 0.0249	0.737 ± 0.006	**0.9481 ± 0.0006**	0.8045 ± 0.0224	0.7325 ± 0.0433
LastFM	0.8011 ± 0.0344	0.7990 ± 0.0444	0.8284 ± 0.0142	0.5196 ± 0.0144	0.9082 ± 0.0015	0.884 ± 0.0	**0.9139 ± 0.0038**
Enron	0.8038 ± 0.0223	0.7120 ± 0.0600	0.8159 ± 0.0231	0.5529 ± 0.015	0.9162 ± 0.0016	0.9051 ± 0.0027	**0.952 ± 0.0057**
SocialEvo	0.8963 ± 0.0228	0.9158 ± 0.0039	0.9244 ± 0.0084	0.6748 ± 0.0021	**0.9298 ± 0.0002**	—	0.896 ± 0.0164
UCI	0.7517 ± 0.0059	0.4297 ± 0.0428	0.8083 ± 0.0237	0.7024 ± 0.0048	0.9177 ± 0.0015	**0.9686 ± 0.0031**	0.9622 ± 0.0167
CollegeMsg	0.5097 ± 0.0306	0.4838 ± 0.0116	0.777 ± 0.0522	0.715 ± 0.0007	0.9163 ± 0.0038	**0.9726 ± 0.0002**	0.9603 ± 0.0173
CanParl	0.5012 ± 0.0155	0.5532 ± 0.0088	0.5727 ± 0.0268	0.5802 ± 0.0069	0.7154 ± 0.0967	**0.8871 ± 0.0139**	0.6214 ± 0.0734
Contact	0.9358 ± 0.0025	0.8650 ± 0.0431	0.952 ± 0.0056	0.5571 ± 0.0047	0.9691 ± 0.0031	**0.9842 ± 0.0**	0.9466 ± 0.0127
Flights	0.9218 ± 0.0094	0.8689 ± 0.0128	0.9519 ± 0.0043	0.8321 ± 0.0041	**0.9834 ± 0.0001**	0.9158 ± 0.0	0.9827 ± 0.003
UNTrade	0.6727 ± 0.0132	0.6467 ± 0.0112	0.5977 ± 0.014	*	**0.7398 ± 0.0007**	0.5915 ± 0.0328	0.6475 ± 0.0664
USLegis	0.5840 ± 0.0129	0.5980 ± 0.0097	0.6128 ± 0.0046	0.5568 ± 0.0078	0.9665 ± 0.0032	**0.9708 ± 0.0009**	0.7453 ± 0.0286
UNVote	0.5121 ± 0.0005	0.4993 ± 0.0103	0.5881 ± 0.0118	0.477 ± 0.0047	0.5911 ± 0.0006	0.586 ± 0.0	**0.779 ± 0.0082**
Taobao	0.701 ± 0.0013	0.7026 ± 0.0006	0.7017 ± 0.0026	0.5261 ± 0.0119	0.7737 ± 0.0027	0.8843 ± 0.0016	**0.9992 ± 0.0002**
Average Rank	4.93	5.4	3.6	6.2	**2**	2.67	2.2
				Inductive New-Old			
Reddit	0.9488 ± 0.0043	0.9549 ± 0.0029	0.9742 ± 0.0004	0.9639 ± 0.0004	0.9848 ± 0.0002	0.9789 ± 0.0017	**0.9949 ± 0.0017**
Wikipedia	0.9084 ± 0.0043	0.8821 ± 0.0031	0.9703 ± 0.0008	0.9178 ± 0.0023	0.9886 ± 0.0002	0.9878 ± 0.0002	**0.9963 ± 0.0021**
MOOC	0.7910 ± 0.0475	0.8274 ± 0.0132	0.8808 ± 0.0326	0.7438 ± 0.0063	**0.949 ± 0.0016**	0.8052 ± 0.0244	0.7487 ± 0.0459
LastFM	0.7305 ± 0.0051	0.6980 ± 0.0364	0.763 ± 0.0231	0.5189 ± 0.003	0.8678 ± 0.0030	0.8311 ± 0.0	**0.9144 ± 0.0013**
Enron	0.7859 ± 0.0134	0.6915 ± 0.0650	0.8100 ± 0.0204	0.5589 ± 0.0235	0.9185 ± 0.003	0.9007 ± 0.0039	**0.9491 ± 0.0079**
SocialEvo	0.8953 ± 0.0303	0.9182 ± 0.0050	**0.9257 ± 0.0086**	0.684 ± 0.0034	0.9155 ± 0.0002	—	0.8793 ± 0.0318
UCI	0.7139 ± 0.0112	0.4259 ± 0.0397	0.8015 ± 0.0269	0.6842 ± 0.0078	0.9176 ± 0.0028	0.9696 ± 0.0039	**0.9748 ± 0.0163**
CollegeMsg	0.5168 ± 0.0360	0.4808 ± 0.0279	0.7725 ± 0.0365	0.7012 ± 0.005	0.9166 ± 0.0032	0.968 ± 0.0018	**0.9725 ± 0.0189**
CanParl	0.5078 ± 0.0005	0.5393 ± 0.0204	0.5691 ± 0.0223	0.5724 ± 0.0063	0.7231 ± 0.085	**0.8847 ± 0.0102**	0.6277 ± 0.0811
Contact	0.9345 ± 0.0027	0.8574 ± 0.0454	0.9527 ± 0.0052	0.556 ± 0.0039	0.9691 ± 0.0028	**0.9841 ± 0.0**	0.9351 ± 0.0202
Flights	0.9172 ± 0.0114	0.8650 ± 0.0126	0.9503 ± 0.0043	0.8285 ± 0.0038	0.9828 ± 0.0002	0.9127 ± 0.0	**0.986 ± 0.0034**
UNTrade	0.6650 ± 0.0106	0.6306 ± 0.0139	0.5959 ± 0.0171	*	**0.7413 ± 0.001**	0.5965 ± 0.0371	0.5812 ± 0.0957
USLegis	0.5801 ± 0.0213	0.5673 ± 0.0098	0.5741 ± 0.0148	0.5596 ± 0.0092	0.9672 ± 0.0029	**0.9682 ± 0.0018**	0.531 ± 0.1
UNVote	0.5208 ± 0.0075	0.5023 ± 0.0204	0.5889 ± 0.0106	0.4787 ± 0.0033	0.5933 ± 0.0007	0.5878 ± 0.0	**0.7789 ± 0.0192**
Taobao	0.6988 ± 0.0024	0.6992 ± 0.0001	0.7025 ± 0.0038	0.5266 ± 0.0239	0.7574 ± 0.0032	0.8617 ± 0.0032	**0.9997 ± 0.0001**
Average Rank	4.87	5.4	3.47	6.13	**1.93**	2.8	2.67
				Inductive New-New			
Reddit	0.9381 ± 0.0090	0.9525 ± 0.0050	0.9811 ± 0.0004	0.9597 ± 0.0043	0.9952 ± 0.0016	0.9875 ± 0.0004	**0.9954 ± 0.0011**
Wikipedia	0.9349 ± 0.0051	0.9261 ± 0.0030	0.9858 ± 0.0007	0.958 ± 0.0039	0.9934 ± 0.0005	0.996 ± 0.0001	**0.9984 ± 0.0008**
MOOC	0.7065 ± 0.0165	0.7217 ± 0.0178	0.8762 ± 0.0038	0.7403 ± 0.0057	**0.9422 ± 0.0003**	0.8048 ± 0.0087	0.6562 ± 0.0287
LastFM	0.8852 ± 0.0090	0.8683 ± 0.0160	0.8754 ± 0.0011	0.5092 ± 0.0333	0.9697 ± 0.0002	0.9628 ± 0.0	**0.9743 ± 0.0015**
Enron	0.6800 ± 0.0017	0.6571 ± 0.0521	0.7644 ± 0.0179	0.531 ± 0.0179	0.9609 ± 0.0051	0.9387 ± 0.0001	**0.9687 ± 0.0051**
SocialEvo	0.6484 ± 0.0491	0.7740 ± 0.0215	0.8791 ± 0.0045	0.4659 ± 0.0068	**0.9318 ± 0.0003**	—	0.9275 ± 0.0471
UCI	0.6393 ± 0.0158	0.4771 ± 0.0100	0.8051 ± 0.021	0.768 ± 0.0041	0.9245 ± 0.0027	**0.9716 ± 0.0016**	0.9472 ± 0.0262
CollegeMsg	0.5320 ± 0.0269	0.5269 ± 0.0049	0.7969 ± 0.0111	0.7832 ± 0.0026	0.9304 ± 0.0024	**0.9762 ± 0.0008**	0.9404 ± 0.0371
CanParl	0.4347 ± 0.0090	0.4430 ± 0.0068	0.5625 ± 0.0396	0.5955 ± 0.0074	0.7005 ± 0.1241	**0.8882 ± 0.0045**	0.5685 ± 0.0326
Contact	0.7531 ± 0.0059	0.6602 ± 0.0395	0.9118 ± 0.0053	0.5449 ± 0.0056	0.9652 ± 0.0012	**0.982 ± 0.0**	0.9495 ± 0.0034
Flights	0.9303 ± 0.0083	0.8900 ± 0.0266	0.9652 ± 0.0022	0.857 ± 0.0056	0.9873 ± 0.0009	0.9411 ± 0.0	**0.9905 ± 0.0014**
UNTrade	0.5922 ± 0.0085	0.5362 ± 0.0147	0.5068 ± 0.0061	*	**0.7458 ± 0.0081**	0.5938 ± 0.0600	0.6876 ± 0.0177
USLegis	0.5390 ± 0.0075	0.5640 ± 0.0192	0.5626 ± 0.0195	0.5324 ± 0.0294	0.9738 ± 0.0058	**0.9787 ± 0.0004**	0.8897 ± 0.0225
UNVote	0.4913 ± 0.0203	0.4728 ± 0.0033	0.5663 ± 0.0093	0.5 ± 0.006	0.5775 ± 0.0022	0.5669 ± 0.0	**0.7198 ± 0.0748**
Taobao	0.7169 ± 0.0013	0.717 ± 0.0011	0.7082 ± 0.001	0.5226 ± 0.0054	0.7847 ± 0.0148	0.9083 ± 0.0013	**0.9996 ± 0.0001**
Average Rank	5.4	5.67	3.93	5.73	**1.73**	2.4	**2.13**
	JODIE	DyRep	TGN	TGAT	CAWN	NeurTW	NAT
Total Rank	4.92	5.39	3.45	6.02	**2.03**	2.69	2.50

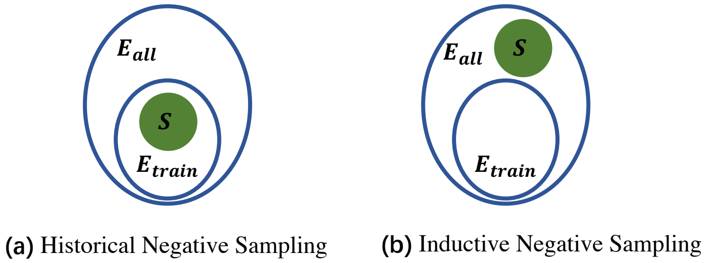
Table 2: ROC AUC results for the node classification task. The top-2 results are highlighted as **bold red** and underlined blue.

Model Dataset	JODIE	DyRep	TGN	TGAT	CAWN	NeurTW	NAT
Reddit	0.6033 ± 0.0173	0.4988 ± 0.0066	0.6216 ± 0.007	0.6252 ± 0.0057	**0.6502 ± 0.0252**	0.6054 ± 0.0352	0.4746 ± 0.0207
Wikipedia	0.8527 ± 0.002	0.8276 ± 0.008	**0.8831 ± 0.0009**	0.8603 ± 0.0051	0.8586 ± 0.0030	0.8470 ± 0.0233	0.5417 ± 0.0474
MOOC	0.6774 ± 0.0066	0.6604 ± 0.0037	0.626 ± 0.0062	0.6705 ± 0.005	0.7271 ± 0.0016	**0.7719 ± 0.0073**	0.5175 ± 0.0143
Average Rank	3.33	5.67	3	2.33	**2**	3.33	7

We will update these *Average Rank* results into the main paper.

2. Negative Sampling

As for NAT and NeurTW can achieved 90%+ or even 95%+ AUC and AP performance on many of these datasets, we have found that BenchTeMP with *Historical Negative Sampling* and *Inductive Negative Sampling* can addressed this issue. [1].



(a) Historical Negative Sampling

(b) Inductive Negative Sampling

Figure 1: *Historical Negative Sampling* and *Inductive Negative Sampling*.

Let E_{all}, E_{train}, S be the set of edges in dataset, the set of edges in train dataset, and the set of negative sampling edges, respectively. *Historical Negative Sampling* and *Inductive Negative Sampling* are illustrated in Figure 1.

- **Historical Negative Sampling**. Sampling negative edges from the set of edges that have been observed during previous timestamps but are absent in the current step, i.e., Sampling negative edges in the E_{train}.
- **Inductive Negative Sampling**. Sampling negative edges in the E_{all} but not observed during training.

We have conducted experiments of NAT on those over-performance datasets (Reddit, Wikipedia, Flights) with *Historical Negative Sampling* and *Inductive Negative Sampling*. Experimental results are shown in Table 3 and Table 4.

Table 3: ROC AUC results of NAT with *Historical Negative Sampling* and *Inductive Negative Sampling* for the dynamic link prediction task.

Sampling	Datasets	Transductive	Inductive	Inductive New-Old	Inductive New-New
Historical	Reddit	0.7759 ± 0.0065	0.8272 ± 0.0036	0.8532 ± 0.0019	0.9097 ± 0.0014
	Wikipedia	0.6992 ± 0.0027	0.7924 ± 0.0022	0.8118 ± 0.0034	0.8448 ± 0.0015
	Flights	0.6145 ± 0.0034	0.6443 ± 0.0276	0.6418 ± 0.0566	0.8019 ± 0.0079
Inductive	Reddit	0.8058 ± 0.0049	0.858 ± 0.0045	0.8746 ± 0.0035	0.9515 ± 0.001
	Wikipedia	0.731 ± 0.0022	0.7609 ± 0.0009	0.7593 ± 0.001	0.8323 ± 0.0045
	Flights	0.6145 ± 0.0034	0.6443 ± 0.0276	0.6418 ± 0.0566	0.8019 ± 0.0079

The experimental results demonstrated the effectiveness of *Historical Negative Sampling* and *Inductive Negative Sampling*.

We leave the exploration of BenchTeMP with *Historical Negative Sampling* and *Inductive Negative Sampling* in detail for future works.

Thank you and best regards!

Table 4: AP results of NAT with *Historical Negative Sampling* and *Inductive Negative Sampling* for the dynamic link prediction task.

Sampling	Datasets	Transductive	Inductive	Inductive New-Old	Inductive New-New
Historical	Reddit	0.7958 ± 0.0097	0.8406 ± 0.0008	0.8558 ± 0.001	0.9063 ± 0.0033
	Wikipedia	0.7128 ± 0.0019	0.7859 ± 0.0022	0.8017 ± 0.0031	0.8267 ± 0.0025
	Flights	0.6287 ± 0.0094	0.6596 ± 0.0248	0.6507 ± 0.0562	0.8428 ± 0.0042
Inductive	Reddit	0.8523 ± 0.0013	0.8902 ± 0.8979	0.8979 ± 0.0035	0.9648 ± 0.0009
	Wikipedia	0.733 ± 0.0025	0.7525 ± 0.0043	0.747 ± 0.0	0.8238 ± 0.0088
	Flights	0.6287 ± 0.0094	0.6596 ± 0.0248	0.6507 ± 0.0562	0.8428 ± 0.0042

References

[1] 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.