Todyformer: Towards Holistic Dynamic Graph Transformers with Structure-Aware Tokenization

Anonymous Author(s)
Affiliation
Address
email

1 Supplementary Material

1.1 Dataset Statistics

In this section, we provide an overview of the statistics pertaining to two distinct sets of datasets utilized for the tasks of Future Link Prediction (FLP) and Dynamic Node Classification (DNC). The initial set, detailed in Table 1, presents information regarding the number of nodes, edges, and unique edges across seven datasets featured in Table **??** and Table **??**. For these three datasets, namely Reddit, Wikipedia, and MOOC, all edge features have been incorporated, where applicable. Furthermore, within this table, the last column represents the percentage of Repetitive Edges, which signifies the proportion of edges that occur more than once within the dynamic graph.

Table 1: Dynamic Graph Datasets. **% Repetitive Edges**: % of edges which appear more than once in the dynamic graph.

Dataset	# Nodes	# Edges	# Unique Edges	Edge Features	Node Labels	Bipartite	% Repetitive Edges
Reddit	11,000	672,447	78,516	✓	✓	✓	54%
Wikipedia	9,227	157,474	18,257	✓	✓	✓	48%
MOOC	7,144	411,749	178,443	✓	✓	✓	53%
LastFM	1980	1,293,103	154,993			✓	68%
UCI	1899	59,835	13838			✓	62%
Enron	184	125,235	2215				92%
SocialEvolution	74	2,099,519	2506				97%

1.1.1 TGB dataset

In this section, we present the characteristics of datasets as proposed by the Dynamic Graph Encoder Leaderboard Huang et al. [2023]. Similar to previous benchmark datasets, we have conducted comparisons regarding the number of nodes, edges, and type of graphs. Additionally, we report the Number of Steps and the Surprise Index, as defined in Poursafaei et al. [2022], which illustrates the ratio of test edges that were not observed during the training phase.

Table 2: Statistics of TGBL Dynamic Graph Datasets

Dataset	# Nodes	# Edges	# Steps	Edge Features	Bipartite	Surprise Index Poursafaei et al. [2022]
Wiki	9,227	157,474	152,757	✓	✓	0.108
Review	352,637	4,873,540	6,865	✓	✓	0.987
Coin	638,486	22,809,486	1,295,720	✓		0.120
Comment	994,790	44,314,507	30,998,030	✓		0.823
Flight	18143	67,169,570	1,385	✓		0.024

Submitted to Temporal Graph Learning Workshop @ NeurIPS 2023, New Orleans. Do not distribute.

1.2 Implementation details

In this section, we elucidate the intricacies of our implementation, providing a comprehensive overview of the specific parameters our model accommodates during hyperparameter optimization. Subsequently, we delve into a discussion of the optimal configurations and setups that yield the best performance for our proposed architecture.

Furthermore, in addition to an in-depth discussion of the baselines incorporated into our paper, we also offer a comprehensive overview of the respective hyperparameter configurations in this section. We are confident that with the open-sourcing of our code upon acceptance and the thorough descriptions of our model and baseline methodologies presented in the paper, our work is fully reproducible.

1.2.1 Evaluation Protocol

Transductive Setup: Under the transductive setting, a dataset is split normally by time, i.e., the model is trained on the first 70% of links, validated on $\%15$ and tested on the rest.

Inductive Setup: In the inductive setting, we strive to test the model's prediction performance on edges with unseen nodes. Therefore, following [Wang et al., 2021], we randomly assign 10% of the nodes to the validation and test sets and remove any interactions involving them in the training set. Additionally, to ensure an inductive setting, we remove any interactions not involving these nodes from the test set.

1.2.2 Best Hyperparameters for Benchmark datasets.

Table 3 displays the hyperparameters that have been subjected to experimentation and tuning for each dataset. For each parameter, a range of values has been tested as follows:

- Window Size (W): This parameter signifies the window length chosen for selecting the input subgraph based on edge timestamps. It falls within the range of $\in$ { 16384, 32768 ,65536, 262144 }.
- Number of Patches: This parameter indicates the count of equal and non-overlapping chunks for each input subgraph. It is the range of $\in \{8, 16, 32\}$.
- #Local Encoders: This parameter represents the number of local encoder layers within each block, and its value falls within the range of $\in \{1, 2\}$.
- Neighbor Sampling (NS) mode: $\in \{uniform, last\}$. In the case of a uniform Neighbor Sampler (NS), it uniformly selects samples from the 1-hop interactions of a given node. Conversely, in last mode, it samples from the most recent interactions.
- Anchor Node Mode: $\in \{GlobalTarget, LocalInput, LocalTarget\}$ depending on the mechanism of neighbor sampling we can sample from nodes within all patches (LocalInput), nodes within the next patch (LocalTarget), or global target nodes (GlobalTarget).
- Batch Size: $\in \{8, 16, 32, 64\}$
- Positional Encoding: $\in \{SineCosine, Time2Vec, Identity, Linear\}$

Dataset	Window Size (W)	Number of Patches	#Local Encoders	NS Mode	Anchor Node Mode	Batch Size
Reddit	262144	32	2	uniform	GlobalTarget	8
Wikipedia	65536	8	2	uniform	GlobalTarget	8
MOOC	65536	8	2	uniform	GlobalTarget	8
LastFM	262144	32	2	uniform	GlobalTarget	8
UCI	65536	8	2	uniform	GlobalTarget	8
Enron	65536	8	2	uniform	GlobalTarget	8
SocialEvolution	65536	8	2	uniform	GlobalTarget	8

Table 3: Best Parameters of the model pipeline after Hyperparameter search

SineCosine is utilized as the Positional Encoding (PE) method following the experiments conducted in Appendix 1.4.1.

Selecting Best Checkpoint: Throughout all experiments, the models undergo training for a duration of 100 epochs, with the best checkpoints selected for testing based on their validation Average Precision (AP) performance.

1.2.3 Best Hyperparameters for TGBL dataset

In this section, we present the optimal hyperparameters used in our architecture design for each TGBL dataset. We conducted hyperparameter tuning for all TGBL datasets; however, due to time constraints, we explored a more limited set of parameters for the large-scale dataset. Despite Todyformer outperforming its counterparts on these datasets, there remains potential for further improvement through an extensive hyperparameter search.

Dataset	Window Size (W)	Number of Patches	First-hop NS size	NS Mode	Anchor Node Mode	Batch Size
TGBWiki	262144	32	256	uniform	GlobalTarget	32
TGBReview	262144	32	64	uniform	GlobalTarget	64
TGBComment	65536	8	64	uniform	GlobalTarget	256
TGBCOin	65536	8	64	uniform	GlobalTarget	96
TGBFlight	65536	8	64	uniform	GlobalTarget	128

Table 4: Optimal Window size W for downstream training.

1.3 More Experimental Result

In this section, we present the additional experiments conducted and provide an analysis of the derived results and conclusions.

1.3.1 FLP result on Benchmark Datasets

Table 5 is an extension of Table **??**, now incorporating the Wikipedia and Reddit datasets. Notably, for these two datasets, Todyformer attains the highest test Average Precision (AP) score in the Transductive setup. However, it secures the second-best position in the Inductive setup for these same datasets.

Table 5: Future link Prediction Performance in AP (Mean ± Std). **Bold** font and ul font represent first- and second-best performance respectively.

Setting	Model	Wikipedia	Reddit	MOOC	LastFM	Enron	UCI	SocialEvol.
Transductive	JODIE	0.956 ± 0.002	0.979 ± 0.001	0.797 ± 0.01	0.691 ± 0.010	0.785 ± 0.020	0.869 ± 0.010	0.847 ± 0.014
	DyRep	0.955 ± 0.004	$0.981 \pm 1e{-4}$	0.840 ± 0.004	0.683 ± 0.033	0.795 ± 0.042	0.524 ± 0.076	0.885 ± 0.004
	TGAT	0.968 ± 0.001	$0.986 \pm 3e{-4}$	0.793 ± 0.006	0.633 ± 0.002	0.637 ± 0.002	0.835 ± 0.003	0.631 ± 0.001
	TGN	0.986 ± 0.001	0.985 ± 0.001	0.911 ± 0.010	0.743 ± 0.030	0.866 ± 0.006	0.843 ± 0.090	0.966 ± 0.001
	CaW	0.976 ± 0.007	$0.988 \pm 2e{-4}$	0.940 ± 0.014	$0.903 \pm 1e{-4}$	0.970 ± 0.001	0.939 ± 0.008	$0.947 \pm 1e{-4}$
	NAT	0.987 ± 0.001	0.991 ± 0.001	0.874 ± 0.004	$0.859 \pm 1e{-4}$	0.924 ± 0.001	0.944 ± 0.002	0.944 ± 0.010
	GraphMixer	0.974 ± 0.001	0.975 ± 0.001	0.835 ± 0.001	0.862 ± 0.003	0.824 ± 0.001	0.932 ± 0.006	$0.935 \pm 3e{-4}$
	Dygformer	0.991 ± 0.0001	0.992 ± 0.0001	0.892 ± 0.005	0.901 ± 0.003	0.926 ± 0.001	0.959 ± 0.001	$0.952 \pm 2e{-4}$
	DyG2Vec	$\underline{0.995 \pm 0.003}$	$\underline{0.996 \pm 2e{-4}}$	$\underline{0.980 \pm 0.002}$	$\underline{0.960 \pm 1e{-4}}$	$\underline{0.991 \pm 0.001}$	$\underline{0.988 \pm 0.007}$	$\underline{0.987 \pm 2e{-4}}$
	Todyformer	$\mathbf{0.996 \pm 2e{-4}}$	$\mathbf{0.998 \pm 8e{-5}}$	$\mathbf{0.992 \pm 7e{-4}}$	$\mathbf{0.976 \pm 3e{-4}}$	$\mathbf{0.995 \pm 6e{-4}}$	$\mathbf{0.994 \pm 4e{-4}}$	$\mathbf{0.992 \pm 1e{-4}}$
Inductive	JODIE	0.891 ± 0.014	0.865 ± 0.021	0.707 ± 0.029	0.865 ± 0.03	0.747 ± 0.041	0.753 ± 0.011	0.791 ± 0.031
	DyRep	0.890 ± 0.002	0.921 ± 0.003	0.723 ± 0.009	0.869 ± 0.015	0.666 ± 0.059	0.437 ± 0.021	$0.904 \pm 3e{-4}$
	TGAT	0.954 ± 0.001	0.979 ± 0.001	0.805 ± 0.006	0.644 ± 0.002	0.693 ± 0.004	0.820 ± 0.005	0.632 ± 0.005
	TGN	0.974 ± 0.001	0.954 ± 0.002	0.855 ± 0.014	0.789 ± 0.050	0.746 ± 0.013	0.791 ± 0.057	0.904 ± 0.023
	CaW	0.977 ± 0.006	$0.984 \pm 2e{-4}$	0.933 ± 0.014	0.890 ± 0.001	0.962 ± 0.001	0.931 ± 0.002	$0.950 \pm 1e{-4}$
	NAT	0.986 ± 0.001	0.986 ± 0.002	$0.832 \pm 1e{-4}$	0.878 ± 0.003	0.949 ± 0.010	0.926 ± 0.010	0.952 ± 0.006
	GraphMixer	$0.966 \pm 2e{-4}$	$0.952 \pm 2e{-4}$	0.814 ± 0.002	0.821 ± 0.004	0.758 ± 0.004	0.911 ± 0.004	$0.918 \pm 6e{-4}$
	Dygformer	$0.985 \pm 3e{-4}$	$\underline{0.988 \pm 2e{-4}}$	0.869 ± 0.004	$0.942 \pm 9e{-4}$	0.897 ± 0.003	0.945 ± 0.001	$0.931 \pm 4e{-4}$
	DyG2Vec	$\mathbf{0.992 \pm 0.001}$	$\mathbf{0.991 \pm 0.002}$	$\underline{0.938 \pm 0.010}$	$\underline{0.979 \pm 0.006}$	$\underline{0.987 \pm 0.004}$	$\underline{0.976 \pm 0.002}$	$\underline{0.978 \pm 0.010}$
	Todyformer	$\underline{0.989 \pm 6e{-4}}$	0.983 ± 0.002	$\mathbf{0.948 \pm 0.009}$	$\mathbf{0.981 \pm 0.005}$	$\mathbf{0.989 \pm 8e{-4}}$	$\mathbf{0.983 \pm 0.002}$	$\mathbf{0.9821 \pm 0.005}$

1.3.2 FLP validation result on TGBL dataset

As discussed in the paper, Todyformer has been compared to baseline methods using the TGBL dataset. Table 6 represents an extension of Table **??** specifically for validation (MRR). The results presented in both tables are in line with counterpart methods outlined in the paper by Huang et al. [2023]. It is important to note that for the larger datasets, TCL, GraphMIxer, and EdgeBank were found to be impractical due to memory constraints, as mentioned in the paper.

1.4 Ablation Studies and Sensitivity Analysis

We conducted an evaluation of the model performance across various parameters and datasets to assess the sensitivity of the major hyperparameters. Figure 1 illustrates the sensitivity analysis regarding the window size and the number of patches, with one parameter remaining constant while the other changes. As highlighted in Xu et al. [2020], recent and frequent interactions display

Table 6: (Validation) Future Link Prediction performance in Validation MRR on TGB Leaderboard datasets.

Model	TGBWiki	TGBReview	TGBCoin	TGBComment	TGBFlight	Avg. Rank ↓
Dyrep	0.411 ± 0.015	0.356 ± 0.016	0.507 ± 0.029	0.291 ± 0.028	0.528 ± 0.022	4.2
TGN	0.737 ± 0.004	$\mathbf{0.465 \pm 0.010}$	0.594 ± 0.023	0.356 ± 0.019	0.739 ± 0.012	2.2
CAWN	$\mathbf{0.794 \pm 0.014}$	0.201 ± 0.002	*OOM*	*OOM*	*OOM*	3
TCL	0.734 ± 0.007	0.194 ± 0.012	*OOM*	*OOM*	*OOM*	5
GraphMixer	0.707 ± 0.014	0.411 ± 0.025	*OOM*	*OOM*	*OOM*	4
EdgeBank	0.641	0.0894	0.1244	0.388	0.492	4.6
Todyformer	0.7821	0.4262	**0.6898**	**0.7434**	**0.7923**	**1.4**

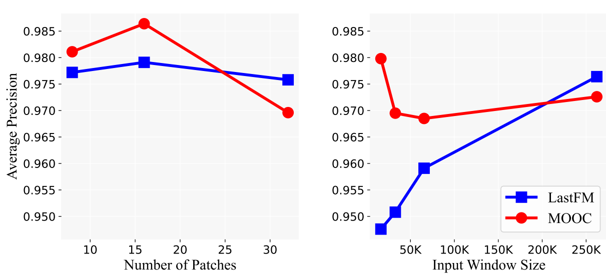

Figure 1: Sensitivity analysis on the number of patches and input window size values on MOOC and LastFM. The plot on the left has a fixed input window size of 262,144, while the one on the right has 32 patches.

Dataset	G. E.	P. E.	Alt. 3	AP
MOOC	✗	✗	✗	0.980
	✓	✗	✗	0.981
	✓	✓	✗	0.987
	✓	✓	✓	**0.992**
LastFM	✗	✗	✗	0.960
	✓	✗	✗	0.961
	✓	✓	✗	0.965
	✓	✓	✓	**0.976**
UCI	✗	✗	✗	0.981
	✓	✗	✗	0.983
	✓	✓	✗	0.987
	✓	✓	✓	**0.993**
SocialEvolution	✗	✗	✗	0.987
	✓	✗	✗	0.987
	✓	✓	✗	0.989
	✓	✓	✓	**0.991**

Table 7: Ablation studies on three major components: global encoder (G. E.), Positional Encoding (P. E.), and number of alternating blocks (Alt. 3)

enhanced predictability of future interactions. This predictability is particularly advantageous for datasets with extensive long-range dependencies, favoring the utilization of larger window size values to capture recurrent patterns. Conversely, in datasets where recent critical interactions reflect importance, excessive emphasis on irrelevant information becomes prominent when employing larger window sizes. Our model, complemented by uniform neighbor sampling, achieves a balanced equilibrium between these contrasting sides of the spectrum. As an example, the right plot in Figure 1 demonstrates that with a fixed number of patches (i.e., 32), an increase in window size leads to a corresponding increase in the validation AP metric on the LastFM dataset. This trend is particularly notable in LastFM, which exhibits pronounced long-range dependencies, in contrast to datasets like MOOC and UCI with medium- to short-range dependencies.

In contrast, in Figure 1 on the left side, with a window size of 262k, we change the number of patches. Specifically, for the MOOC dataset, performance exhibits an upward trajectory with an increase in the number of patches from 8 to 16; however, it experiences a pronounced decline when the number of patches reaches 32. This observation aligns with the inherent nature of MOOC datasets, characterized by their relatively high density and reduced prevalence of long-range dependencies. Conversely, when considering LastFM data, the model maintains consistently high performance even at 32 patches. In essence, this phenomenon underscores the model's resilience on datasets featuring extensive long-range dependencies, illustrating a trade-off between encoding local and contextual features by tweaking the number of patches.

In table 7, we conduct ablation studies on the major design choices of the encoding network to assess the roles of the three hyperparameters separately: a) Global encoder, b) Alternating mode c) Positional Encoding. Across the four datasets, the alternating approach exhibits significant performance variation compared to others, ensuring the mitigation of over-smoothing and the capturing of long-range dependencies. The outcomes of the single-layer vanilla transformer as global encoder attain the second-best position, affirming the efficacy of our global encoder in enhancing expressiveness. Finally, the global encoder without PE closely resembles the model with only a local encoder (e.i. DyG2Vec MPNN model).

Positional Encoding	Anchor_Node_Mode	Average Precision ↑
SineCosinePos	global target	0.9901
Time2VecPos	global target	0.989
IdentityPos	global target	0.99
LinearPos	global target	0.9886
SineCosinePos	local input	0.9448

Table 8: **Ablation Study on Positional Encoding Options on MOOC Dataset:** This table compares the validation performance at the same epoch across various setups.

Table 9: Sensitivity analysis on number of patches and target window size

dataset	Input Window size	Number of Patches	Average Precision ↑
LastFM	262144	8	0.9772
LastFM	262144	16	0.9791
LastFM	262144	32	0.9758
MOOC	262144	8	0.9811
MOOC	262144	16	0.9864
MOOC	262144	32	0.9696
LastFM	16384	32	0.9476
LastFM	32768	32	0.9508
LastFM	65536	32	0.9591
LastFM	262144	32	0.9764
MOOC	16384	32	0.9798
MOOC	32768	32	0.9695
MOOC	65536	32	0.9685
MOOC	262144	32	0.9726

1.4.1 Complementary Sensitivity Analysis and Ablation Study

In this section, we have presented the specifics of sensitivity and ablation experiments, which, while of lesser significance in our hyper-tuning mechanism, contribute valuable insights. In all tables, the Average Precision scores reported in the table are extracted from the same epoch on the validation set. Table 9 showcases the influence of varying input window sizes and patch sizes on two datasets. Table 8 illustrates the effects of various PEs, including SineCosine, Time2VecKazemi et al. [2019], Identity, Linear, and a configuration utilizing Local Input as the Anchor Node Mode. The table presents a comparison of results for these different PEs. Notably, the architecture appears to be relatively insensitive to the type of PE used, as the results all fall within a similar range. However, it is worth mentioning that SineCosine PE slightly outperforms the others. Consequently, SineCosine PE will be selected as the primary module for all subsequent experiments.

References

Shenyang Huang, Farimah Poursafaei, Jacob Danovitch, Matthias Fey, Weihua Hu, Emanuele Rossi, Jure Leskovec, Michael Bronstein, Guillaume Rabusseau, and Reihaneh Rabbany. Temporal graph benchmark for machine learning on temporal graphs. *arXiv preprint arXiv:2307.01026*, 2023.

Farimah Poursafaei, Shenyang Huang, Kellin Pelrine, and Reihaneh Rabbany. Towards better evaluation for dynamic link prediction. *Advances in Neural Information Processing Systems*, 35:32928–32941, 2022.

Yanbang Wang, Yen-Yu Chang, Yunyu Liu, Jure Leskovec, and Pan Li. Inductive representation learning in temporal networks via causal anonymous walks. In *Proc. Int. Conf. on Learning Representations*, 2021.

Da Xu, Chuanwei Ruan, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. Inductive representation learning on temporal graphs. *Proc. Int. Conf. on Representation Learning*, 2020.

Seyed Mehran Kazemi, Rishab Goel, Sepehr Eghbali, Janahan Ramanan, Jaspreet Sahota, Sanjay Thakur, Stella Wu, Cathal Smyth, Pascal Poupart, and Marcus Brubaker. Time2vec: Learning a vector representation of time. *arXiv preprint arXiv:1907.05321*, 2019.