BENCHTEMP: A General Benchmark for Evaluating Temporal Graph Neural Networks

Authors' Response to All Reviewers

Main paper (MP): https://openreview.net/pdf?id=rnZm2vQq31

Appendix (APP): https://openreview.net/attachment?id=rnZm2vQq31&name=supplementary_material

Dear reviewers:

We sincerely appreciate all your feedback and valuable comments! We have made dedicated efforts to improve our paper quality according to your valuable comments and suggestion, respectively.

In this work, we conduct a comprehensive benchmark termed BENCHTEMP on the state-of-the-art TGNN models.

The major contributions of this work are summarized below.

1. **(Sec. 1 in MP)** We present BENCHTEMP, a general benchmark for evaluating temporal graph neural network (TGNN) models over a wide range of tasks and settings. We release the datasets, code, and leaderboard.
    - Datasets - https://zenodo.org/record/8267846.
    - Code - https://github.com/qianghuangwhu/benchtemp.
    - BENCHTEMP Leaderboards - https://my-website-6gnpiaym0891702b-1257259254.tcloudbaseapp.com/.
2. - **(Sec. 3.1 in MP, Sec. F in APP)** We collect/construct**21** benchmark temporal graph datasets with a unified preprocess to ensure dataset consistency. We have made engineering efforts to unify "Node Feature Initialization" and "Node Reindexing".
    - **(Sec. F in APP)** In particular, we have included four datasets (**eBay-Large, DGraphFin, YouTubeReddit-Large, Taobao-Large**), with up to several million edges and nodes.
    - **(Sec. F in APP)** Besides, we are working on sharing the ***eBay-Small*** and ***eBay-Large*** datasets in a way that ensures availability and justifies the research purpose. eBay provide a Google form for the applicants: https://forms.gle/bP1RmyVJ1C6pgyS66 (the applicants can remain anonymous).
3. **(Sec. 3.2 in MP)** We proposed a unified benchmark pipeline. In this way, we standardize the entire lifecycle of benchmarking TGNNs.
    - Pipeline
        - Dynamic *link prediction* task pipeline:
          Dataset –> DataLoader –> EdgeSampler -> Model -> EarlyStopMonitor -> Evaluator -> Leaderboard.
        - Dynamic *node classification* task pipeline:
          Dataset –> DataLoader –> Model –> EarlyStopMonitor –> Evaluator –> Leaderboard.

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

4. **(Sec. 4 in MP, Sec. D and Sec. F in APP)** We extensively compare seven representative TGNN models on the benchmark datasets, regarding different tasks, settings, metrics (*AUC, AP, Average Rank*) , and efficiency (*Runtime, RAM, GPU, Inference time*)(New)). Note that *Average Rank* and *Inference Time* are two **new** metrics.
   - **(Sec. 4.2 in MP, Sec. F.1.1 in APP)** Dynamic *link prediction* task on **21** temporal graph datasets
     - Diverse settings: Transductive, Inductive, Inductive New-Old, Inductive New-New.
     - Prediction performance and mode efficiency.
   - **(Sec. 4.3 in MP, Sec. F.1.2 in APP)** Dynamic *node classification* task on **6** temporal graph datasets
     - **(Sec. 4.3 in MP** Implementation of the dynamic node classification task on five datasets (Reddit, Wikipedia, MOOC, eBay-Small, and eBay-Large) with binary node labels (label: 0 and 1).
     - **(Sec. G in APP)** For the **first** time in TGNNs, we evaluate the dynamic ***multi-class*** node classification task on the ***large-scale*** DGraph dataset with multiple node labels (label: 0, 1, 2, and 3)
     - Comparison of the model efficiency on dynamic node classification task.
   - **(Sec. 4 in MP, Sec. F in APP)** Evaluation of both model performances and efficiency on the newly added ***large-scale*** datasets.
5. **(Sec. 4 in MP, Sec. F in APP)** We thoroughly discuss the empirical results and draw insights for future studies on TGNNs..
   - **(Sec. 4 in MP, Sec. F in APP)** Experimental results reveal that ***NeurTW performs poorly on efficiency and the joint-neighborhood operation of NAT does not perform well on the node classification task compared to its superior performance on the link prediction task.***
   - **(Sec. 4 in MP, Sec. F in APP)** Memory-based TGNNs (JODIE, DyRep, and TGN) are unfit for temporal graphs *with a large amount of nodes.*
   - **(Sec. H in APP)** We further conduct ***ablation studies*** to verify the effectiveness of neural ordinary differential equations (*NODEs*) of NeurTW on datasets with a large time granularity and time intervals.
   - **(Sec. I in APP)** The strategy of ***random subgraph sampling*** with a constant number of edges demonstrates demonstrate that the effectiveness of the CAWN based on temporal walk changes in response to changes of graph density.
   - **(Sec. J in APP)** Furthermore, we have discussed BENCHTEMP with ***Historical Negative Sampling*** and ***Inductive Negative Sampling*** and leave it for future work.

Thank you and best regards!

Yours sincerely,

Authors