A Dataset Checklist

1. Submission introducing new datasets must include the following in the supplementary materials
   (a) Dataset documentation and intended uses. [Yes] We include the datasheets for datasets of TGB 2.0 in Appendix I.
   (b) URL to website/platform where the dataset/benchmark can be viewed and downloaded by the reviewers. [Yes] The website link and documentation link is included in Appendix D.
   (c) URL to Croissant metadata record documenting the dataset/benchmark available for viewing and downloading by the reviewers. [Yes] The croissant metadata record link is `https://object-arbutus.cloud.computecanada.ca/tgb/tgb2_croissant.json`.
   (d) Author statement that they bear all responsibility in case of violation of rights, etc., and confirmation of the data license. [Yes] Yes, we bear all responsibility and also state this in Appendix E.
   (e) Hosting, licensing, and maintenance plan. [Yes] Yes, we discuss the hosting and licensing plan in Appendix D.
2. To ensure accessibility, the supplementary materials for datasets must include the following:
   (a) Links to access the dataset and its metadata. [Yes] Yes, all links are provided in Appendix E and D.
   (b) The dataset itself should ideally use an open and widely used data format. Provide a detailed explanation on how the dataset can be read. For simulation environments, use existing frameworks or explain how they can be used. [Yes] The dataset is automatically downloaded and processed by the TGB 2.0 code and presented in ML ready format.
   (c) Long-term preservation: It must be clear that the dataset will be available for a long time, either by uploading to a data repository or by explaining how the authors themselves will ensure this. [Yes] TGB 2.0 datasets are maintained via Digital Research Alliance of Canada (funded by the Government of Canada).
   (d) Explicit license: Authors must choose a license, ideally a CC license for datasets, or an open source license for code (e.g. RL environments). [Yes] Yes, all dataset licenses are provided in Appendix E. The TGB 2.0 code is provided in the MIT license.
   (e) Add structured metadata to a dataset's meta-data page using Web standards (like schema.org and DCAT): This allows it to be discovered and organized by anyone. If you use an existing data repository, this is often done automatically. [Yes] We provide the croissant metadata record, the link is `https://object-arbutus.cloud.computecanada.ca/tgb/tgb2_croissant.json`.
   (f) Highly recommended: a persistent dereferenceable identifier (e.g. a DOI minted by a data repository or a prefix on identifiers.org) for datasets, or a code repository (e.g. GitHub, GitLab,...) for code. If this is not possible or useful, please explain why. [Yes] The DOI for the project is `https://zenodo.org/doi/10.5281/zenodo.11480521`.

B Limitations

This work exclusively considers the continuous-time setting for THG datasets. Depending on the application, either the continuous-time or discrete-time setting may be more appropriate. However, the continuous-time setting is often regarded as the more general framework. Nonetheless, many THG methods are designed for discrete settings. Thus, as future work, discretized versions of the datasets for comparative analysis between discrete methods could be added.

Additionally, the TGB 2.0 dataset collection currently includes datasets from only five distinct domains. Notably, domains such as biological networks and citation networks are not represented. To

address this limitation, we plan to expand the dataset collection by incorporating additional datasets based on community feedback, thereby enhancing the diversity and comprehensiveness of the dataset repository.

C Broader Impact

Impact on Temporal Graph Learning. Recently, the availability of large graph benchmarks accelerates research in the field [25, 24, 10]. By providing a standardized benchmarking framework, TGB 2.0 will accelerate the development and evaluation of new models for temporal knowledge graphs and temporal heterogeneous graphs. Researchers can build upon a common foundation, leading to more rapid and robust advancements in this field. In addition, the introduction of a unified evaluation framework addresses reproducibility issues, which are critical for scientific progress. The comprehensive evaluation facilitated by TGB 2.0 ensures that new methods are rigorously tested against state-of-the-art baselines, leading to more robust and well-validated models. This contributes to higher standards in research and more reliable outcomes. Overall, this work has the potential to significantly impact both the academic research community and practical applications, driving forward the understanding and utilization of multi-relational temporal graphs in various fields.

Potential Negative Impact. The TGB 2.0 datasets may limit the utilization and mining of other TG datasets. If the datasets are not representative of the broader set of real-world data, this could lead to biased or unfair outcomes when models are applied in practice. Similarly, the community might become overly dependent on the TGB 2.0 framework, potentially hindering the exploration of alternative benchmarking methodologies or the development of diverse evaluation protocols that might be more suitable for specific contexts or emerging subfields. Moreover, when the focus is mainly on quantitative performance metrics, it might overshadow the importance of qualitative assessments and other critical factors such as interpretability, fairness, and ethical considerations in model development and deployment. To avoid this issue, we plan to update TGB regularly with community feedback as well as adding additional datasets and tasks.

D Dataset Documentation and Intended Use

All datasets presented by TGB 2.0 are intended for academic use and their corresponding licenses are listed in Appendix E. We also anonymized the datasets, to remove any personally identifiable information where appropriate. For the ease of access, we provide the following links to the TGB 2.0 benchmark suits and datasets.

- The code is available publicly on TGB2 Github: `https://github.com/JuliaGast/TGB2`. The code will also be merged into TGB Github.
- Dataset and project documentations can be found at: `https://tgb.complexdatalab.com/`.
- Tutorials and API references can be found at: `https://docs.tgb.complexdatalab.com/`.
- Hugging face link for main dataset files is `https://huggingface.co/datasets/andrewsleader/TGB/tree/main`.
- ML croissant metadata file link is `https://object-arbutus.cloud.computecanada.ca/tgb/tgb2_croissant.json`.

Maintenance Plan. We plan to continue to improve and develop TGB 2.0 based on community feedback to provide a reproducible, open and robust benchmark for temporal multi-relational graphs. We will maintain and improve the TGB 2.0, TGB and TGB-Baselines github repository, while the TGB 2.0 datasets are maintained via Digital Research Alliance of Canada (funded by the Government of Canada).

E Dataset Licenses and Download Links

In this section, we present dataset licenses and the download link (embedded in dataset name). The datasets are maintained via Digital Research Alliance of Canada funded by the Government of Canada. As authors, we confirm the data licenses as indicated below and that we bear all responsibility in case of violation of rights. We also included the metadata for datasets in the ML croissant format [2]. The ML croissant metadata link is `https://object-arbutus.cloud.computecanada.ca/tgb/tgb2_croissant.json`.

- `tkgl-smallpedia`: Wikidata License. See license information from Wikidata License Page. Property and lexeme namespaces is made available under the Creative Commons CC0 License. Text in other namespaces is made available under the Creative Commons Attribution-ShareAlike License. Here is the data source link.
- `tkgl-polecat`: CC0 1.0 DEED license. Here is the data source link.
- `tkgl-icews`: Custom Dataset License. The detailed license information can be found here. Restrictions on use: these materials are subject to copyright protection and may only be used and copied for research and educational purposes. The materials may not be used or copied for any commercial purposes. Here is the data source link.
- `tkgl-wikidata`: Wikidata License. See license information from Wikidata License Page. Property and lexeme namespaces is made available under the Creative Commons CC0 License. Text in other namespaces is made available under the Creative Commons Attribution-ShareAlike License. Here is the data source link.
- `thgl-software`: CC-BY-4.0 license. This dataset is curated from GH Arxiv code which has the MIT License. Content based on www.gharchive.org is released under the CC-BY-4.0 license. To avoid any personal identifiable information, we anonymized all nodes to integers. The raw data can be found here.
- `thgl-forum`: CC BY-NC 2.0 DEED license. The raw data source is here [51].
- `thgl-myket`: CC BY-NC 4.0 DEED license. A smaller subset of this dataset is available on Github.
- `thgl-github`: CC-BY-4.0 license.This dataset is curated from GH Arxiv code which has the MIT License. Content based on www.gharchive.org is released under the CC-BY-4.0 license. To avoid any personal identifiable information, we anonymized all nodes to integers. The raw data can be found here.

F Dataset Statistics

Figure 5 shows how the number of edges change over time for TKG datasets. Figures 6 shows how the number of edges change over time for THG datasets. While most datasets exhibit fluctuations in the number of edges around a constant level, `tkgl-wikidata` stands out with a significant upward trend in the number of edges over the years, indicating a surge in events, particularly in recent years. In addition, noteworthy deviations in timesteps are apparent. TKG datasets display anomalous timesteps characterized by minimal edge numbers, particularly evident during the Covid pandemic for `tkgl-icews`. Conversely, for the THG datasets the occurrence of zero-edge timesteps is not indicative of outliers; rather, it reflects the continuous nature of the data, where not every second entails an event occurrence. THG datasets exhibit instances of exceptionally high edge counts per timestep, such as in the case of `thgl-forum` with up to 120 edges per timestamp.

Figure 7 shows the top ten most frequent edge types in TKG datasets. Figure 8 shows the top ten most frequent edge types in THG datasets. Note that TKG datasets in general has more edge types than THG datasets. Most common THG relations usually share similar portion of edges in the dataset while TKG relations shares different portion of edges.

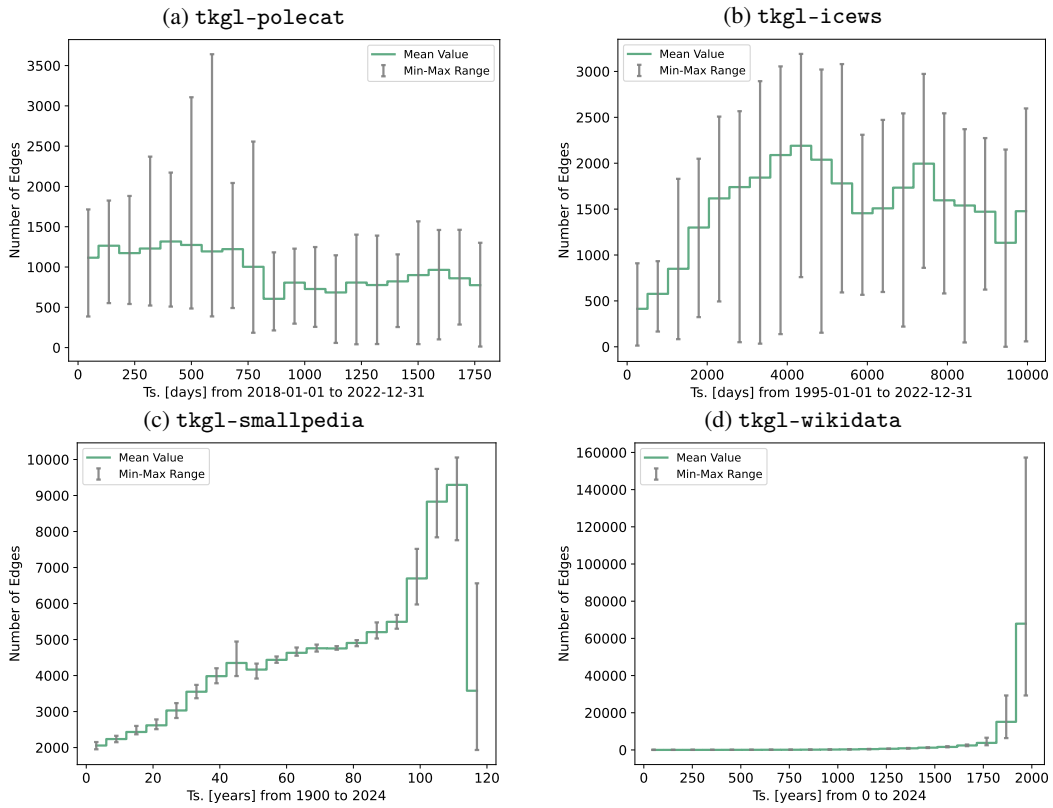

Figure 5: Dataset Edges over time for TKG.

G Experimental Details

In the following, we provide additional experimental details such as the computing resources, resource consumption, hyperparameters, and runtime statistics.

G.1 Computing Resources

We ran all experiments on either Narval or Béluga cluster of Digital Research Alliance of Canada or the Mila, Québec AI Institute cluster. For the experiments on the Narval cluster, we ran each experiment on a Nvidia A100 (40G memory) GPU with 4 CPU nodes (from either of the AMD Rome 7532 @ 2.40 GHz 256M cache L3, AMD Rome 7502 @ 2.50 GHz 128M cache L3, or AMD Milan 7413 @ 2.65 GHz 128M cache L3 available type) each with 100GB memory. For experiments on the Béluga cluster, we rand each experiments on a NVidia V100SXM2 (16G memory) GPU wiht 4 CPU nodes (from Intel Gold 6148 Skylake @ 2.4 GHz) each with 100GB memory. For the experiments on the Mila cluster, we ran each experiment on an RTX8000 (40G memory) GPU or an V100 (32G memory) GPU with 4 CPU nodes (from either of the AMD Rome 7532 @ 2.40 GHz 256M cache L3, AMD Rome 7502 @ 2.50 GHz 128M cache L3, or AMD Milan 7413 @ 2.65 GHz 128M cache L3 available type). The upper limit of RAM was set to 1056GB.

A seven-day time limit was considered for each experiment. For all non deterministic methods, i.e. all methods besides Edgebank and the Recurrency Baseline, we repeated each experiments five times and reported the average and standard deviation of different runs. It is noteworthy that except for the reported baseline results, the other models, all evaluated by their original source code, throw an out of memory error or do not finish in the given time limit for the medium and large datasets on all available resources includiing Narval, Béluga, and Mila clusters.

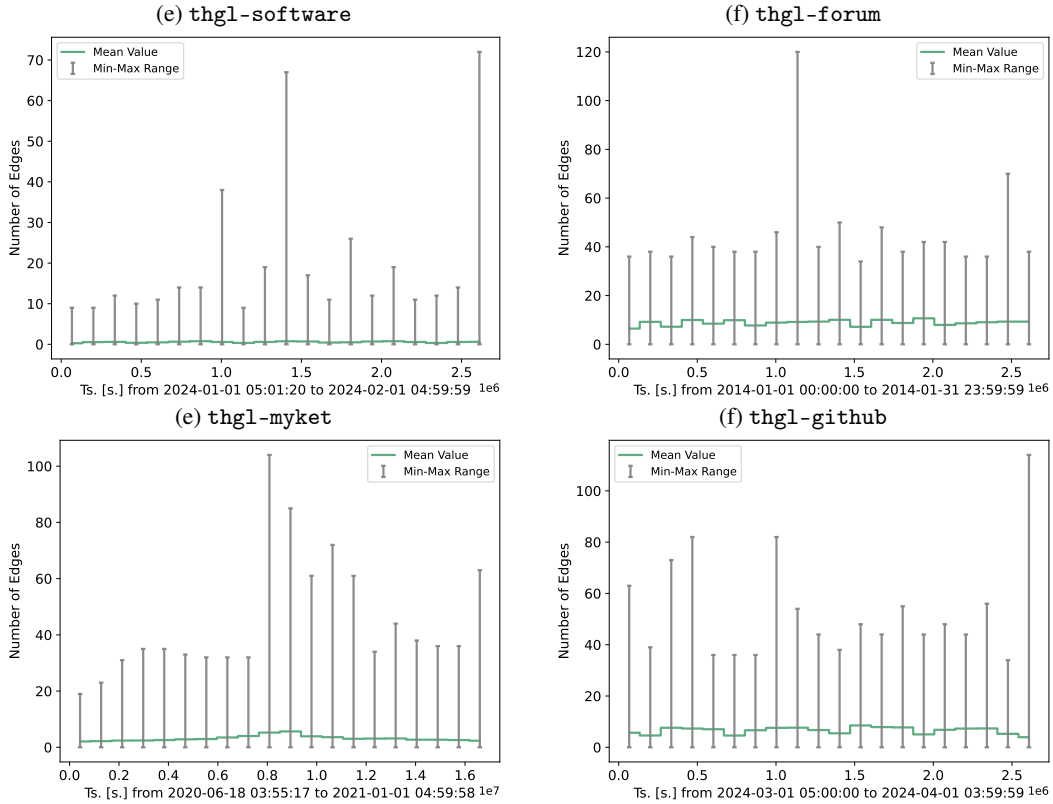

Figure 6: Dataset Edges over time for THG.

Table 4: GPU memory usage in **GB** for the *Temporal Knowledge Graph Link Prediction* task for the methods that run on GPU. We report the average across 5 runs.

Method	`tkgl-smallpedia`	`tkgl-polecat`	`tkgl-icews`	`tkgl-wikidata`
RE-GCN [39]	20.9	21.2	24.3	OOM
CEN [37]	28.8	41.0	31.6	OOM

G.2 GPU Usage Comparison

In Table 4 and 5, we report the average GPU usage of TKG and THG methods on the dataset across 5 trials. Note that the Recurrency Baseline, EdgeBank, and TLogic only require CPU thus no GPU usage is reported. For TKG, some methods such as CEN on `tkgl-polecat` have higher GPU usage when compared to others. For THG, scalability is a significant issue, as most methods involve high GPU usage and often result in out-of-memory errors, especially with larger datasets. Although STHN maintains manageable GPU usage, it requires substantial RAM to compute the subgraphs, making it impractical for use in all environments.

G.3 Runtime Comparison

In Table 6 and Table 7 we report the inference times as well as the total time for training, validation and testing for each method for TKG and THG experiments. For the non-deterministic methods, we report the average across 5 runs. The tables illustrate that both, inference times, as well as total times vary significantly across methods.

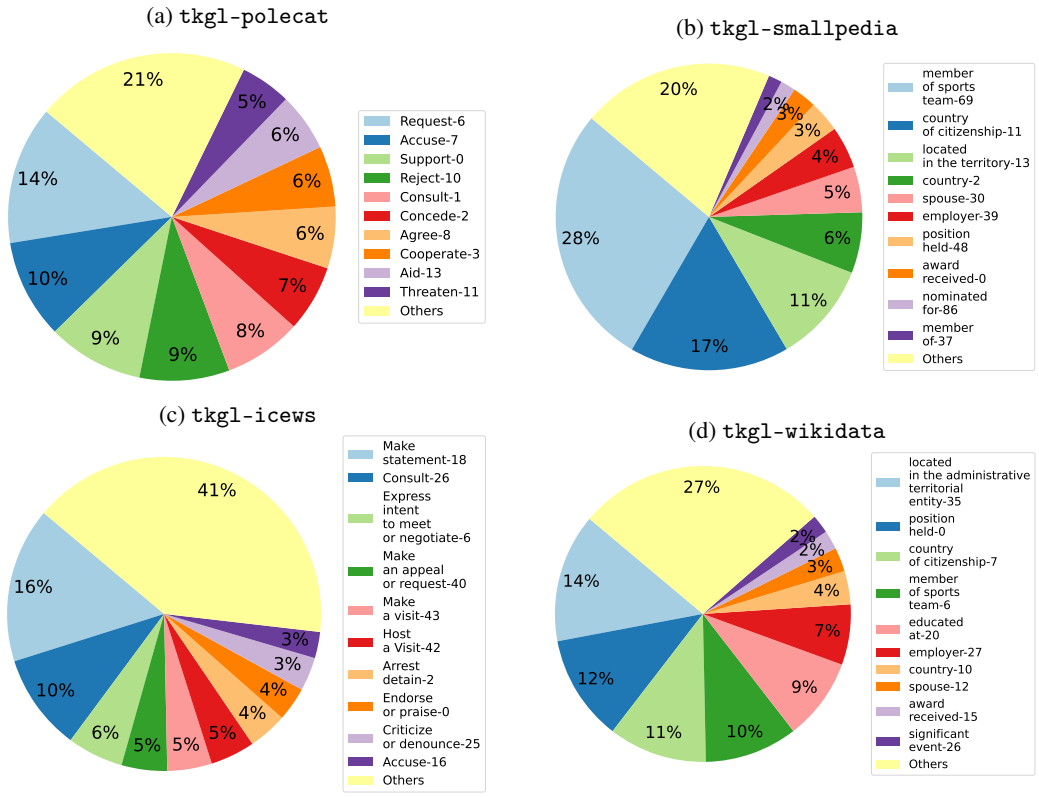

Figure 7: Edge type ratios in TGB 2.0 TKGs. We include the 10 most frequent edge types.

Table 5: GPU memory usage in **GB** for *Temporal Heterogeneous Graph Link Prediction* task. We report the average across 5 runs.

Method	`thgl-software`	`thgl-forum`	`thgl-myket`	`thgl-github`
TGN [58]	7	8	-	-
$\text{TGN}_{\text{edge-type}}$	10	12	-	-
STHN [36]	15	-	-	-

G.3.1 Hyperparameters

If not stated otherwise, for each method we use the hyperparameter setting as reported in the original papers, please see Table 8. Whereas further hyperparameter tuning could further improve performance of each method, it was out of scope for this work. We only change the hyperparameter values only if the methods would not finish with the given time or memory limit. In this case, we follow recommendations from [14] (to decrease rule length and window size for TLogic), from the authors of [13] (to decrease the window length for the Recurrency Baseline), and from the authors of [39] (to decrease the history length for RE-GCN and CEN).

G.4 Experimental Observations

Several methods encountered memory limitations or did not complete within the designated time constraints. Thus, as described in Section 5, their performance is not reported. In the following, we provide additional details on the problems of individual methods:

- RE-GCN and CEN run out of GPU memory for `tkgl-wikidata`, even if severely limiting embedding dimension and history length.

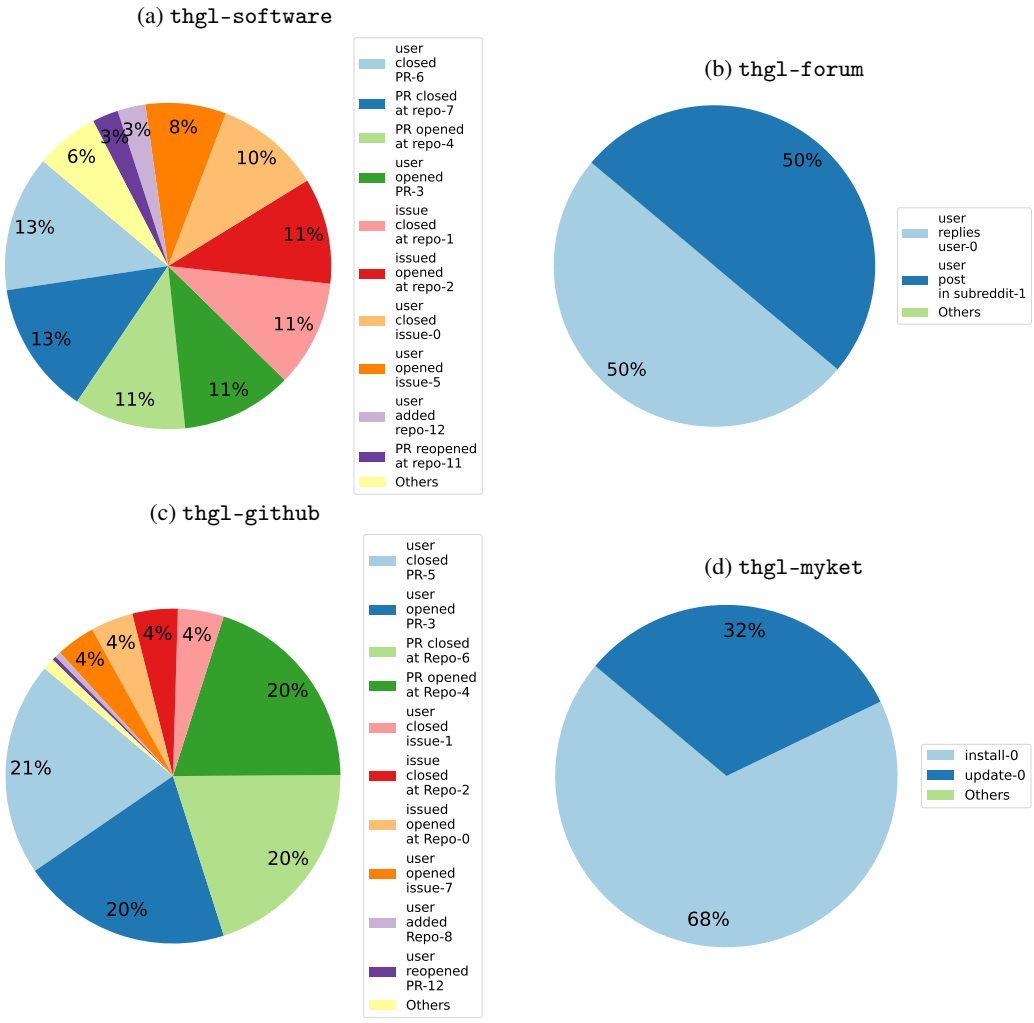

Figure 8: Edge type ration in TGB 2.0 THGs.

Table 6: Inference time as well as total train and validation times for *Temporal Knowledge Graph Link Prediction* task in **seconds**. For non-deterministic methods, we report the average across 5 different runs.

Method	tkgl-smallpedia		tkgl-polecat		tkgl-icews		tkgl-wikidata	
	Test	Total	Test	Total	Test	Total	Test	Total
EdgeBank$_{tw}$ [54]	2,935	5,810	46,629	94,475	311,278	600,929	5,445	8,875
EdgeBank$_{\infty}$ [54]	4,417	8,259	31,713	64,157	203,268	412,774	4,814	7,923
RecurrencyBaseline$_{train}$ [13]	310	9,895	4,500	8,343	-	-	-	-
RecurrencyBaseline$_{default}$ [13]	316	659	3,392	80,378	11,756	30,110	-	-
RE-GCN [39]	165	3,895	1,766	45,877	6,848	114,370	-	-
CEN [37]	331	14,493	2,726	77,953	8,999	202,477	-	-
TLogic [44]	331	803	75,654	138,636	60,413	128,391	-	-

- Recurrency Baseline does not finish in the designated time constraint for the large THG datasets `thgl-myket` and `thgl-github` and the large TKG dataset `tkgl-wikidata`.

- TLogic does not finish in the designated time constraint for `tkgl-wikidata`. Further, we reduced the rule length to 1 to fit in the time constraint and memory limitations for the introduced datasets.

Table 7: Inference time as well as total train and validation time for *Temporal Heterogeneous Graph Link Prediction* task in **seconds**. For the non-deterministic methods, we report the average across 5 different runs.

Method	thgl-software		thgl-forum		thgl-myket		thgl-github	
	Test	Total	Test	Total	Test	Total	Test	Total
$\text{EdgeBank}_{\text{tw}}$ [54]	102	203	1,158	2,329	4,820	9,603	295	301
EdgeBank_{∞} [54]	107	212	1,148	2,303	4,956	10,017	282	296
$\text{RecurrencyBaseline}_{\text{default}}$ [13]	62,259	114,124	32,539	65,114	-	-	-	-
TGN [58]	686	66,290	7,654	8,8659	-	-	-	-
$\text{TGN}_{\text{edge-type}}$	567	39,427	8,241	111,494	-	-	-	-
STHN [36]	52,101	102,943	-	-	-	-	-	-

Table 8: Hyperparameter choices. Values that are different from the original papers are **bolded**. In case we modify the values for different datasets, we report so in the respective columns.

Method	Hyperparameter Values	
	All Datasets	Dataset-specific
TLogic	**rule_lengths = 1**, window $= 0$, top_k $= 20$	`tkgl-icews`: window $= 500$
RE-GCN	n_hidden $= 200$, n_layers $= 2$, dropout $= 0.2$, lr $= 0.001$, n_bases $= 100$, train_history_len $= 3$, test_history_len $= 3$	
CEN	n_hidden $= 200$, n_layers $= 2$, dropout $= 0.2$, lr $= 0.001$, n_bases $= 100$, n_layers $= 2$, **train_history_len** $= 3$ test_history_len $= 3$, **start_history_len** $= 2$, dilate_len $= 1$	
RecB	$\lambda = 0.1$, $\alpha = 0.99$, window $= 0$	`tkgl-icews`: window $= 500$
	All Datasets	Dataset-specific
RecB	$\lambda = 0.1$, $\alpha = 0.99$, window $= 0$	
TGN	lr $= 1e-04$, mem_dim $= 100$, time_dim $= 100$, emb_dim $= 100$, num_neighbors $= 10$	
$\text{TGN}_{\text{edge-type}}$	lr $= 1e-04$, mem_dim $= 100$, time_dim $= 100$, emb_dim $= 100$, num_neighbors $= 10$, edge_emb_dim $= 16$	
STHN	lr $= 5e-04$, max_edges $= 50$, window_size $= 5$, dropout $= 0.1$, time_dims $= 100$, hidden_dims $= 100$	

- STHN model has very high memory consumption, requires 185 GB of RAM on the small `thgl-software` dataset (mostly due to subgraph computations). On the rest of THG datasets, it rans out of memory.

- TGN and $\text{TGN}_{\text{edge-type}}$ run out of GPU memory for both `thgl-myket` and `thgl-github`, even if limiting embedding dimension to time_dim $=$ mem_dim $=$ emb_dim $= 16$ and edgeType_dim $= 16$.

G.5 Ablation Study on Negative Sample Generation

Here, we compare results for evaluation on the full set of nodes (*1-vs-all*) versus a limited number of negative samples q (*1-vs-q*). We also compare our sampling method based on destination nodes of each edge type (*1-vs-q* (ours)) with that of random sampling (*1-vs-q* (random)). We select the `tkgl-smallpedia` dataset and report results for the Recurrency Baseline as well as Edgebank, as both methods perform competitively while being deterministic. Table 9 confirms expectations: random negative sampling yields the highest MRR values. MRR values for our destination-aware negative sampling demonstrate a closer proximity to the full sampling (*1-vs-all*) for both methodologies. Notably, employing the 1-vs-all approach yields the lowest MRR for both test and validation sets, underscoring the importance of comprehensive evaluations whenever feasible. However, particularly evident in the case of Edgebank, the adoption of negative sampling significantly reduces test time, changing from approximately 3000 seconds to 70 seconds.

Table 9: MRR and Runtime for Edgebank and the Recurrency Baseline (RecB) on the `tkgl-smallpedia` dataset for three different strategies for Negative Sample Generation.

Strategy	Method	MRR		Runtime [s.]	
		valid	test	test	total
1-vs-1000 (random)	$\text{RecB}_{\text{default}}$ [13]	0.755	0.734	278	692
	$\text{EdgeBank}_{\text{tw}}$ [54]	0.706	0.576	72	141
1-vs-1000 (ours)	$\text{RecB}_{\text{default}}$ [13]	0.642	0.608	282	703
	$\text{EdgeBank}_{\text{tw}}$ [54]	0.612	0.495	104	210
1-vs-all	$\text{RecB}_{\text{default}}$ [13]	0.640	0.570	316	659
	$\text{EdgeBank}_{\text{tw}}$ [54]	0.457	0.353	2935	5810

Table 10: Number of Edges and timestamps for train, validation and test set for each dataset in TGB 2.0.

	Temporal Knowledge Graphs (`tkgl-`)				Temporal Heterogeneous Graphs (`thgl-`)			
Dataset	`smallpedia`	`polecat`	`icews`	`wikidata`	`software`	`forum`	`github`	`myket`
# Train Quadruples	387,757	1,246,556	10,861,600	6,982,503	1,042,866	16,630,396	12,249,711	37,542,951
# Valid Quadruples	81,033	266,736	2,326,157	1,434,950	223,469	3,563,658	2,624,934	8,044,922
# Test Quadruples	81,586	266,318	2,325,689	1,438,750	223,471	3,563,653	2,624,932	804,4915
# All Quadruples	550,376	1,779,610	15,513,446	9,856,203	1,489,806	23,757,707	17,499,577	53,632,788
# Train Timesteps	98	1,193	7,187	1,999	485,863	1,805,376	1,703,696	9,935,183
# Valid Timesteps	10	329	1,341	12	99,500	393,000	382,882	2,274,936
# Test Timesteps	17	304	1,696	14	104,186	360,081	423,837	2,617,971
# All Timesteps	125	1,826	10,224	2,025	689,549	2,558,457	2,510,415	14,828,090

G.6 Detailed information on Train, Validation, and Test Splits

As described in Section 4, we split all datasets chronologically into the training, validation, and test sets, respectively containing 70%, 15%, and 15% of all edges. Because we ensure that edges for a timesteps can only be in either train or validation or test set, and because the number of edges over time are not constant, the cuts are not strict. We provide more details on the exact splits in Table 10.

H More Details on Methods

In the following we will describe the methods that we selected for our experiments.

H.1 Temporal Knowledge Graph Forecasting

For our experiments we select methods from a variety of methods from the previous literature. We base our selection on a) code availability, b) comparatively high performance in previous studies on smaller datasets (following results as reported in [14] and [13], i.e. we exclude methods that are reported to have lower MRRs on all previous datasets as compared to the Recurrency Baseline), and c) we exclude methods that have reported to have long runtimes or high GPU memory consumption on the existing smaller datasets (e.g. [20] for the GDELT dataset [14]). This results in the following TKG baselines:

- *RE-GCN [39]* learns from the sequence of Knowledge Graph snapshots recurrently by combining a convolutional graph Neural Network with a sequential Neural Network model. It also incorporates a static graph constraint to include additional information like entity types.
- *CEN [37]* integrates a GCN capable of handling evolutional patterns of different lengths through a learning strategy that progresses from short to long patterns. This model can adapt to changes in evolutional patterns over time in an online setting, being updated with historical facts during testing.

- *TLogic [44]* is a symbolic framework that learns temporal logic rules via temporal random walks, traversing edges backward in time through the graph. It applies these rules to events preceding the query, considering both the confidence of the rules and the time differences for scoring answer candidates.
- Recurrency Baseline [13] is a baseline method that predicts recurring facts by combining scores based on strict recurrency, considering the recency and frequency of these facts, and scores based on relaxed recurrency, which accounts for the recurrence of parts of the query. Two versions of this baseline are tested: $\text{RecB}_{\text{default}}$, which uses default parameter values, and $\text{RecB}_{\text{train}}$, which selects parameter values based on a grid search considering performance on the validation set.

H.2 Temporal Heterogeneous Graph Forecasting

- *TGN [58]* represents a comprehensive framework designed for learning on dynamic graphs in continuous time. Its components include a memory module, message function, message aggregator, memory updater, and embedding module. During testing, TGN updates the memories of nodes with edges that have been newly observed. Additionally, to incorporate edge types into the TGN, we devised a variant of the TGN capable of utilizing edge type information. This was achieved by generating embeddings from the edge types, which were then concatenated with the original messages within the TGN model.
- *STHN [36]* designed for continuous-time link prediction on Temporal heterogeneous networks that efficiently manages dynamic interactions. The architecture consists of a *Heterogeneous Link Encoder* with type and time encoding components, which embed historical interactions to produce a temporal link representation. The process continues with *Semantic Patches Fusion*, where sequential representations are divided into different patches treated as token inputs for the Encoder, and average mean pooling compresses these into a single vector. Finally, the framework combines the representations of nodes u and v, utilizing a fully connected layer and *CrossEntropy* loss for link prediction, effectively capturing complex temporal information and long-term dependencies.

I Datasheets for Datasets

This section answers questions about this work based on Datasheets for Datasets [15].

I.0.1 Motivation

- **For what purpose was the dataset created?** Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description. TGB 2.0 is curated for realistic, reproducible and robust evaluation for temporal multi-relational graphs. Specifically there are four TKG datasets and four THG datasets, all designed for the dynamic link property prediction task.
- **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?** `thgl-software` and `thgl-github` datasets are based on Github data collected by GH Arxiv. `thgl-forum` dataset is derived from user and subreddit interactions on Reddit. `thgl-myket` dataset was generated by the data team of the Myket Android application market. `tkgl-smallpedia` and `tkgl-wikidata` datasets are constructed from the Wikidata Knowledge Graph. `tkgl-polecat` is based on the POLitical Event Classification, Attributes, and Types (POLECAT) dataset. `tkgl-icews` is extracted from the ICEWS Coded Event Data. Detailed Dataset information is found in Section 4.
- **Who funded the creation of the dataset?** If there is an associated grant, please provide the name of the grantor and the grant name and number. Funding information is provided in Acknowledgement Section.

I.0.2 Composition

- **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?** Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

  The datasets primarily consist of nodes and edges in graph structures, representing various entities and their interactions:

  - `thgl-software` **and** `thgl-github`**:** Nodes represent entities like users, pull requests, issues, and repositories. Edges indicate interactions among these entities.
  - `thgl-forum`**:** Comprises user and subreddit nodes with edges for user replies and posts.
  - `thgl-myket`**:** Features nodes as users and Android applications, with edges detailing install and update interactions. These datasets facilitate tasks like predicting future interactions or activities, utilizing a graph model to depict relationships in various domains such as software development, online communities, and socio-political contexts.
  - `tkgl-smallpedia` **and** `tkgl-wikidata`**:** Includes Wikidata entities as nodes with edges as temporal and static relations.
  - `tkgl-polecat` **and** `tkgl-icews`**:** Focus on socio-political actors as nodes with edges representing coded interactions.

- **How many instances are there in total (of each type, if appropriate)?** The detailed dataset statistics can be found in Section 4, Table 1.

- **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?** If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

  The datasets are curated from the raw source. In some cases, some data filtering is done to remove low degree nodes. More details on dataset curation is found in Section 4. For `thgl-myket`, the data provider first focused on users interacting with the platform within a two-week period and randomly sampled 1/3 of the users. The install and update interactions for these users were then tracked for three months before and after the two-week period.

- **What data does each instance consist of?** "Raw" data (e.g., unprocessed text or images)or features? In either case, please provide a description.

  The data contains the multi-relational temporal graph structure in the form of csv files as well as pre-generated negative samples for reproducible evaluation.

- **Is there a label or target associated with each instance?** If so, please provide a description.

  We focus on the dynamic link property prediction (or link prediction) task thus the goal is to predict edges in the graph in the future. Therefore, no specific task labels are provided. We also provide both node and edge type information for THGs and edge type information for TKGs.

- **Is any information missing from individual instances?** If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

  No, we provide information required for ML on temporal graphs.

- **Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?** If so, please describe how these relationships are made explicit.

  The dataset themselves are classfied into TKG or THG datasets, specified by the prefix `tkgl` or `thgl`. The relations between nodes are assigned with an edge type which is provided in the csv file.

- **Are there recommended data splits (e.g., training, development/validation, testing)?** If so, please provide a description of these splits, explaining the rationale behind them.

  Yes, the recommended split uses a 70/15/15 split, and the data is split chronologically. Pleasee see Table 10 for details on the dataset splits.

- **Are there any errors, sources of noise, or redundancies in the dataset?** If so, please provide a description.

  No. However, datasets such as `tkgl-smallpedia` and `tkgl-wikidata` are extracted from Wikipedia where the knowledge is crowd-sourced, and thus may contain errors.

- **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?** If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

  The dataset is self-contained.

- **Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' nonpublic communications)?** If so, please provide a description.

  No, all data are gathered from public sources and we have anonymized user information where appropriate.

- **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?** If so, please describe why.

  No.

- **Does the dataset identify any subpopulations (e.g., by age, gender)?** If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

  No.

- **Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?** If so, please describe how.

  No, we have anonmyzied users' information where appropriate.

- **Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?** If so, please provide a description.

  No.

I.0.3 Collection Process

- **How was the data associated with each instance acquired?** Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

  The data is extracted from online public data sources. The data described different relations between entities. The data sources are found in Appendix E and dataset details are in Section 4.

- **What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)?** How were these mechanisms or procedures validated? Software APIs.

  The datasets are curated via Python scripts written by authors, these can be found on the project Github.

- **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**

  For `thgl-myket`, the users where selected randomly among the users that have interactions with the platform in a two-week period. For `tkgl-smallpedia`, `tkgl-wikidata`, the dataset was filtered by Wiki page ID. `thgl-software` and `thgl-github`, nodes with low degrees are filtered out.

- **Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?**

  Datasets are obtained from public online sources. For `thgl-myket` dataset, the interaction record of users of the platform were collected, anonymized without any personal identifiers, the data collection is discussed in the applications' privacy document. No crowdworkers are involved.

- **Over what timeframe was the data collected?** Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

  Dataset timeframe and details are in Section 4.

- **Were any ethical review processes conducted (e.g., by an institutional review board)?** If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

  No.

- **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?**

  All datasets are obtained via websites except for `thgl-myket` which were provided by the the Myket Android application market team. Links to data sources are in Appendix E.

- **Were the individuals in question notified about the data collection?** If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

  All datasets are curated from existing sources except `thgl-myket`. The data collection was discussed in the applications' privacy document.

- **Did the individuals in question consent to the collection and use of their data?** If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

  We use public data sources where data is already collected. The data collection was discussed in the applications' privacy document.

- **If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?** If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

  [N/A]

- **Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?** If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

No, however the datasets are for temporal graph research purposes only, they are used to benchmark existing methods and have been anonymized appropriately.

I.0.4 Preprocessing/cleaning/labeling

- **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?** If so, please provide a description. If not, you may skip the remaining questions in this section. No.
- **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?** If so, please provide a link or other access point to the "raw" data.
  [N/A]
- **Is the software that was used to preprocess/clean/label the data available?** If so, please provide a link or other access point.
  [N/A]

I.0.5 Uses

- **Has the dataset been used for any tasks already?** If so, please provide a description.
  Yes, all datasets have been tested and benchmarked in this work, see Section 5.
- **Is there a repository that links to any or all papers or systems that use the dataset?** If so, please provide a link or other access point.
  Yes, all paper references are provided in this paper. All data sources are discussed in Appendix E.
- **What (other) tasks could the dataset be used for?**
  The THG datasets can be used for other tasks such as user churn prediction and more. The TKG datasets can be used to study how knowledge changes over time.
- **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?** For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?
  No, the datasets are for benchmarking purposes only and for researchers.
- **Are there tasks for which the dataset should not be used?** If so, please provide a description.
  No and we discuss potential negative impacts in Appendix C.

I.0.6 Distribution

- **Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?** If so, please provide a description.
  The dataset is released to the public for benchmarking on TKGs and THGs.
- **How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?** Does the dataset have a digital object identifier (DOI)?
  Yes, the DOI for the project is `https://zenodo.org/records/11480522` (will point to all future version as well). The dataset download links are provided in Appendix E. TGB 2.0 datasets are maintained via Digital Research Alliance of Canada (funded by the Government of Canada).
- **When will the dataset be distributed?** The dataset is already publicly available.

- **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?** If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions. The dataset licenses are listed in Appendix E.
- **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?** If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions. All license terms are discussed in Appendix E.
- **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?** If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation. No.

I.0.7 Maintenance

- **Who will be supporting/hosting/maintaining the dataset?** TGB 2.0 datasets are maintained via Digital Research Alliance of Canada (funded by the Government of Canada).
- **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?** The curator of the dataset (Shenyang Huang) can be contacted via email: `shenyang.huang@mail.mcgill.ca`
- **Is there an erratum?** If so, please provide a link or other access point. No
- **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?** If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (e.g., mailing list, GitHub)? Yes, the datasets will be updated based on community feedback, mainly via the main TGB Github issues.
- **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)?** If so, please describe these limits and explain how they will be enforced. No.
- **Will older versions of the dataset continue to be supported/hosted/maintained?** If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers. Any new dataset version will be annouced on Github and the TGB website.
- **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?** If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.

  Yes, first they can reach out by email to `shenyang.huang@mail.mcgill.ca` or raise a Github issue.