TGB 2.0: A Benchmark for Learning on Temporal Knowledge Graphs and Heterogeneous Graphs

Julia Gastinger*, **Shenyang Huang***, Mikhail Galkin, Erfan Loghmani, Ali Parviz, Farimah Poursafaei, Jacob Danovitch, Emanuele Rossi, Ioannis Koutis, Heiner Stuckenschmidt, Reihaneh Rabbany, Guillaume Rabusseau

Website, Paper, Github, Pypi, Documentation
https://tgb.complexdatalab.com/

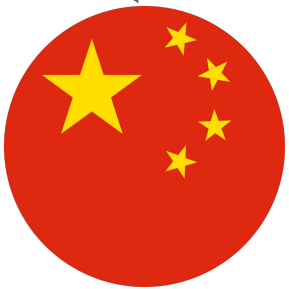
Real World Relations Evolve Over Time

Donald Trump

is president of

visits

United States of America



People's Republic of China

November 2017

Joe Biden

is president of

visits

United States of America

United Kingdom

June 2021

Multi-Relational Temporal Graphs

Temporal Knowledge Graph (TKG)

is a set of quadruples (s, r, o, t)

- subject s and object o $\in$ V
- relation r $\in$ R and timestamp t.

❑ Knowledge bases

❑ Political event networks

Temporal Heterogeneous Graph (THG)

is a set of quadruples (s, r, o, t)

- With node type function $\phi: V \rightarrow A$.

❑ Software networks

❑ Social networks

❑ Interaction networks

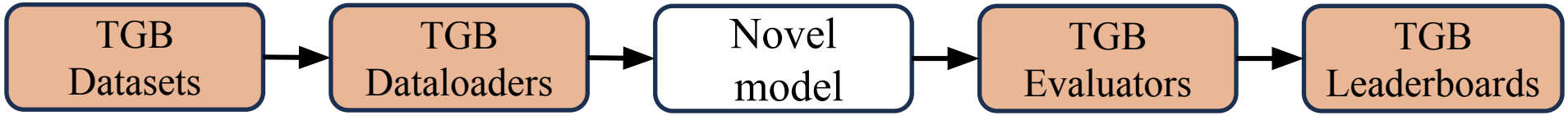
➢ **Task: Temporal Graph Extrapolation (Link Prediction)**

- Predict links between nodes in future time steps
- For a given query, e.g. (s, r, ?, t+), rank all nodes using a scoring function

Limitations in Existing Literature

- Inconsistent Evaluation
  - TKG evaluation has inconsistent metrics, settings and dataset versions
  - THG evaluation often only has a single random negative per positive edge
- Limited Dataset Size
  - Common TKG and THG datasets have < 2 million edges , <1 million nodes

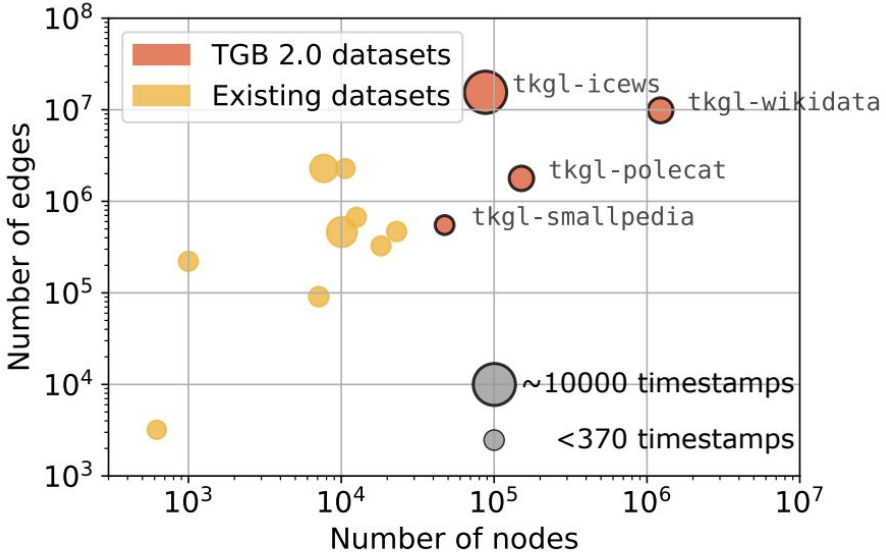
TGB 2.0



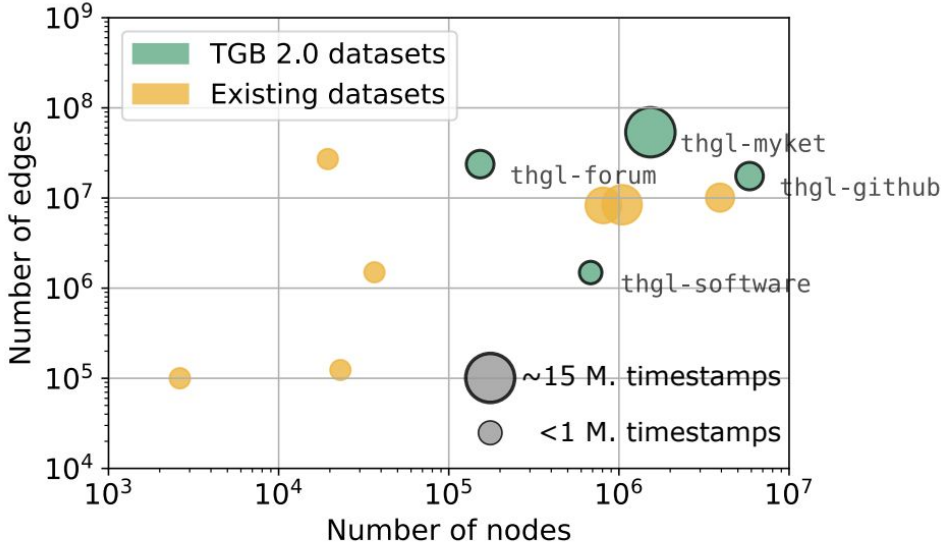
- **Large and Diverse Datasets**
  - Four TKG and four THG datasets from five domains
- **Automatic Data processing and Loading**
  - Processed into numpy, PyTorch and PyG formats
- **Reproducible Evaluation**
  - Data loaders, evaluators provided
- **Public and Online Leaderboard**
  - Open for community submissions

https://tgb.complexdatalab.com/

TGB 2.0 Datasets



(a) Novel Temporal Knowledge Graphs



(b) Novel Temporal Heterogeneous Graphs

➢ Orders of magnitude larger in # edges, # nodes, # timestamps.

Evaluation Protocol

- **Task: temporal graph extrapolation** (link prediction)
- **Metric: time-aware filtered MRR**, rank true target out of many negatives
  - Select # of negative edges based on tradeoff between evaluation completeness & efficiency

TKG Evaluation	THG Evaluation
Predict (s, r, ?, t+) & (?, r, o, t+) - **1-vs-all:** for smaller datasets, sample all ns samples. - **1-vs-q:** sample **q** negatives with same edge type as true edge.	Predict (s, r, ?, t+) - **1-vs-q**: for all THG datasets, sample **q** negatives with same node type as the true destination.

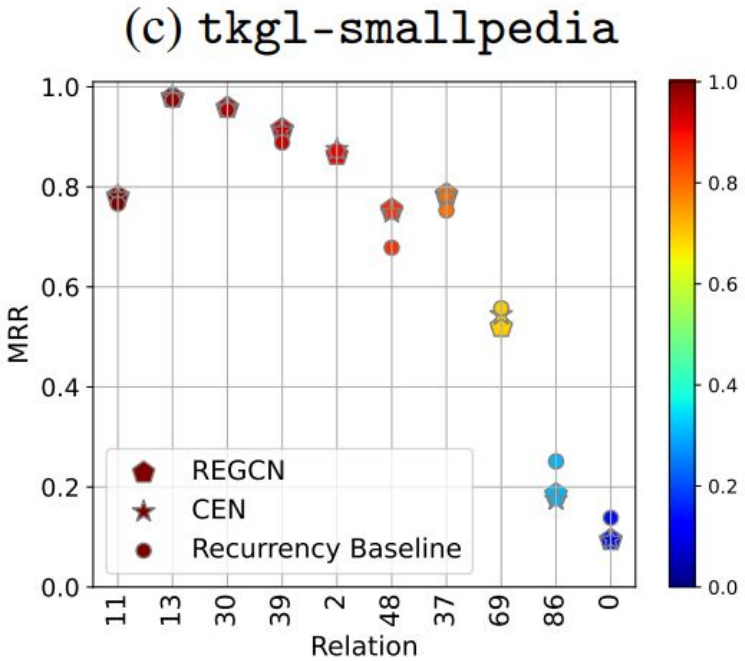
TKG Experiments

Table 2: **MRR** results for *Temporal Knowledge Graph Link Prediction* task. We report the average and standard deviation across 5 runs. First place is **bolded**, second place underlined.

Method	tkgl-smallpedia		tkgl-polecat		tkgl-icews		tkgl-wikidata	
	Validation	Test	Validation	Test	Validation	Test	Validation	Test
$\text{EdgeBank}_{\text{tw}}$ [57]	0.457	0.353	0.058	0.056	0.020	0.020	0.633	0.535
EdgeBank_{∞} [57]	0.401	0.333	0.048	0.045	0.008	0.009	0.632	0.535
$\text{RecB}_{\text{train}}$ [15]	0.639	0.605	0.203	0.198	0.270	**0.211**	OOT	OOT
$\text{RecB}_{\text{default}}$ [15]	0.542	0.486	0.170	0.167	0.264	0.206	OOT	OOT
RE-GCN [41]	0.631±0.001	0.594±0.001	0.191±0.003	0.175±0.002	0.232±0.003	0.182±0.003	OOM	OOM
CEN [39]	**0.646**±0.001	**0.612**±0.001	0.204±0.002	0.184±0.002	0.244±0.002	0.187±0.003	OOM	OOM
TLogic [47]	0.631±0.000	0.595±0.001	**0.236**±0.001	**0.228**±0.001	**0.287**±0.001	0.186±0.001	OOT	OOT

- The heuristic recurrency baseline performs competitively
- Scalability of existing methods are limited
- Out of Memory (OOM) / Out of Time (OOT)

Recurring Relations are Easier to Predict



- Warmer -> more recurring
- More recurring relations have higher MRR across methods

THG Experiments

Table 3: **MRR** results for *Temporal Heterogeneous Graph Link Prediction* task. We report the average and standard deviation across 5 runs. First place is **bolded**, second place underlined.

Method	thgl-software		thgl-forum		thgl-github		thgl-myket	
	Validation	Test	Validation	Test	Validation	Test	Validation	Test
$\text{EdgeBank}_{\text{tw}}$ [57]	0.279	0.288	0.534	0.534	0.355	0.374	0.248	0.245
EdgeBank_{∞} [57]	0.399	0.449	0.612	0.617	0.403	0.413	0.430	0.456
$\text{RecB}_{\text{default}}$ [15]	0.106	0.099	0.552	0.561	OOT	OOT	OOT	OOT
TGN [61]	0.299±0.012	0.324±0.017	0.598±0.086	0.649±0.097	OOM	OOM	OOM	OOM
$\text{TGN}_{\text{edge-type}}$	0.376±0.010	0.424±0.013	**0.767**±0.005	**0.729**±0.009	OOM	OOM	OOM	OOM
STHN [38]	**0.764**±0.025	**0.731**±0.005	OOM	OOM	OOM	OOM	OOM	OOM

- Models that use edge type / node type information perform well
- STHN is SOTA on software but least scalable
- Out of Memory (OOM) / Out of Time (OOT)

- Website: https://tgb.complexdatalab.com/
- Documentation: https://docs.tgb.complexdatalab.com/
- Github: https://github.com/shenyangHuang/TGB
- `pip install py-tgb`
- Welcome to submit to our leaderboard.
- Contact: shenyang.huang@mail.mcgill.ca