Supplementary Material

1 OUR KG BENCHMARK

Table 1: Our Benchmark Statistics. The number of nodes and edges (RDF triples) is in millions. The number of node types (n-type) and edge types (e-type) is tens to thousands.

KG-Dataset	#nodes	#edges	#n-type	#e-type
MAG-42M	42.4M	166M	58	62
YAGO-30M	30.7M	400M	104	98
DBLP-15M	15.6M	252M	42	48
ogbl-wikikg2	2.5M	17M	9.3K	535
YAGO3-10	123K	1.1M	23	37

1.1 An Overview: KG datasets and GNN tasks

Our KG datasets include real KGs, such as MAG, DBLP, YAGO and WikiKG. The datasets are extracted from real KGs of different application domains: general-fact KGs (Yago [1] and ogbl-wikikg2 [2]) and academic KGs (MAG[3] and DBLP [4]). These KGs contain up to 42.4 million vertices, 400 million triples, and tens to thousands of node/edge types. Existing HGNN methods for LP tasks on large KGs require excessive computing resources. Hence, we also used YAGO3-10 [5], a smaller version of YAGO, and ogbl-wikikg2 [4], a dataset extracted from Wikidata. It contains 2.5M entities and 535 edge types. Tables 1 and 2 summarize the details of the used KGs and our defined NC and LP tasks, respectively. Our benchmark will be available for further study.

Our benchmark includes NC and LP tasks. In KGs, an NC task is categorized into single- or multi-label classifications, as defined in **??**. Our benchmark followed existing benchmarks [4, 6, 7] and defined single-label NC tasks. We choose the accuracy metric for these NC tasks to evaluate the performance. The train-valid-test splits are either split using a logical predicate that depends on the task or stratified random split with 80% for training, 10% for validation and 10% for testing. Table 2 summarizes the details of the split schema and ratio. Our benchmark includes two NC tasks per KG.

For LP tasks, KGs contain many edge types that vary in importance to a specific task. For example, a user from an academic background will be interested in the predicates (edge type) *discipline* and *affiliation* in Wikidata but not in predicates related to movies and films. Moreover, training HGNNs for an LP task on a large KG of many edge types is computationally expensive. Hence, we define our LP tasks for a specific predicate to enable the HGNN methods to train the models with our available computing resources: a Linux machine with 32 cores and 3TB RAM. We choose the Hits@10 metric to evaluate the task performance following SOTA methods [2, 4, 7]. The train-valid-test splits are either split using three versions of the KG based on time or randomly. The details of the LP tasks are summarized in Table 2.

Table 2: A summary of our GNN tasks: Task Types (TT) are single-label node classification (NC) or missing entity link prediction (LP).

TT	Name	KG	Split	Ratio	Metric
NC	PV	MAG-42M	Time	84/9/7	Accuracy
NC	PD	MAG-42M	Time	87/8/5	Accuracy
NC	PC	YAGO-30M	Random	80/10/10	Accuracy
NC	CG	YAGO-30M	Random	80/10/10	Accuracy
NC	PV	DBLP-15M	Time	79/10/11	Accuracy
NC	AC	DBLP-15M	Time	80/10/10	Accuracy
LP	AA	DBLP-15M	Time	99/0.7/0.3	Hits@10
LP	PO	ogbl-wikikg2	Time	94/2.5/3.5	Hits@10
LP	CA	YAGO3-10	Random	99/0.5/0.5	Hits@10

[1] YAGO-4: https://yago-knowledge.org/downloads/yago-4
[2] ogbl-wikikg2: https://ogb.stanford.edu/docs/linkprop/#ogbl-wikikg2
[3] MAG-2020-05-29: https://makg.org/rdf-dumps/
[4] DBLP versions March 2022 and Jan, March, and May 2023: https://dblp.org/rdf/release/
[5] YAGO3-10: https://paperswithcode.com/dataset/yago

1.2 Node Classification Datasets:

MAG-42M Paper-Venue (PV): The dataset extracted from The MAG KG [3] to train a GNN model that predicts the publication venue of a paper. The target nodes are 736K papers, and the prediction labels are their associated venues. The set of paper is the same set in OGBN-MAG [4] dataset while including all neighbours nodes for a depth of 3 hops. The number of classes (distinct venues) is 349. The training set includes papers published until 2018, the validation set includes papers published in 2019, and the test set includes papers published in 2020. The number of node types is 58, and the number of edge types is 62. The number of edges is 166M, and the number of nodes is 42M, as shown in table 1.
DBLP-15M Paper-Venue (PV): The dataset extracted from DBLP [1] to train a model that predicts the publication venue of a paper. For the set of labels (distinct venues), we choose the top 50 frequent venues. Any paper that does not belong to the top 50 is excluded. This task includes 1.2M papers. The training set includes papers published before 2020, the validation set includes papers published in 2020, and the test set includes papers published in 2021.
MAG-42M Paper-Discipline (PD): The dataset extracted from MAG [3] to train a model that predicts the discipline of a paper based on its authors and citations. The target nodes are 710K papers, and the prediction labels are their associated disciplines. The set of paper is the same set in OGBN-MAG [4] dataset while including all neighbours nodes for a depth of 3 hops. For the label set (distinct disciplines), we choose the top 50 frequent disciplines. We exclude any paper that does not belong to this set. The splitting criteria are the same as MAG-42M Paper-Venue (PV) dataset.
DBLP-15M Author Affiliation-Country (AC): The dataset extracted from the DBLP KG[1] to train a model that predicts the affiliation country of an author based on its collaboration history. The target vertices are 92K authors with type *(person)*. The label set includes the top 50 frequent countries. The author nodes are randomly split into train-valid-test sets.
Yago4-30M KG (PlaceCountry (PC)): We used YAGO4 [5] to train a model that predicts the country of a place based on its neighbour places. There are 734K target vertices of type *(place)*. The label set includes the top 200 frequent countries. The places nodes are randomly split into train-valid-test sets.
Yago-4 KG (Creative-work Genre (CG)): We also used YAGO4 [5] to train a model that predicts the genre of creative work. There are 128K target vertices of type *(Creative-Work)*. The label set includes the top 50 frequent genres. The creative work nodes are randomly split into train-valid-test sets.

1.3 Link Prediction Benchmark datasets:

Yag03-10 Connected Airports (CA): We defined an LP task on YAGO3 [**?**] to predict the missing triples $\langle v_t, p, ? \rangle$, where v_t is of type airport and p is the *isConnected* predicate. The *isConnected* triples in the original validation and testing subsets are used for the validation and testing of the model.
ogbl-Wikikg2 Person Occupations (PO): The ogbl-wikikg2 dataset is used for an LP task that predicts the missing occupation links for a person (edge type $P106$ in Wikidata). The dataset is split according to time with three versions of May, August, and November 2015 for training, validation and testing, respectively.
DBLP-15M Author PrimaryAffiliations (AA): We defined another LP task on DBLP [1] to predict the missing triples $\langle v_t, p, ? \rangle$, where v_t is of type author and p is the *PrimaryAffiliation* predicate. Our training, validation, and test sets included the DBLP versions of Jan, March, and May 2023, respectively.

2 NODE CLASSIFICATION RESULTS

Figures 1,2, and 3 show the results for the six node classification tasks.

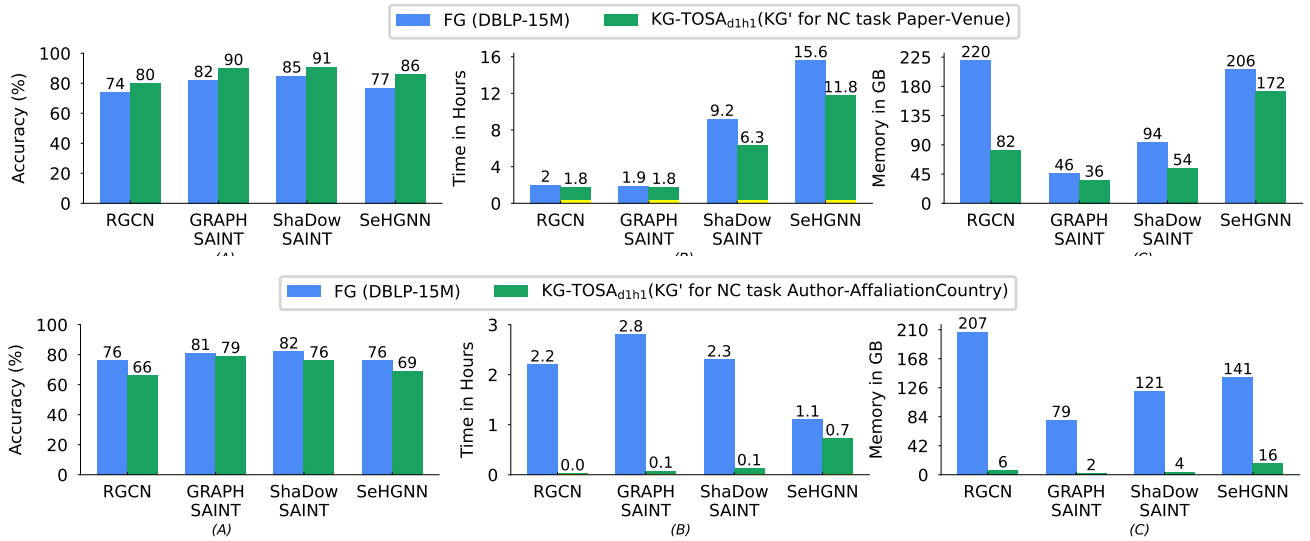

Figure 1: Performance of RGCN, Graph-SAINT, ShaDow-SAINT, and SeHGNN in the NC tasks. (A) Accuracy (higher is better), (B) Training-Time (lower is better), (C) Training-Memory (lower is better). The figures on top show the results for the paper-venue classification task on DBLP. The figures at the bottom show the results for the Author-Affalition_Country classification task. KG-TOSA enables all methods to reduce memory and time while improving accuracy or keeping comparable scores, even with KG-TOSA's preprocessing time in yellow.

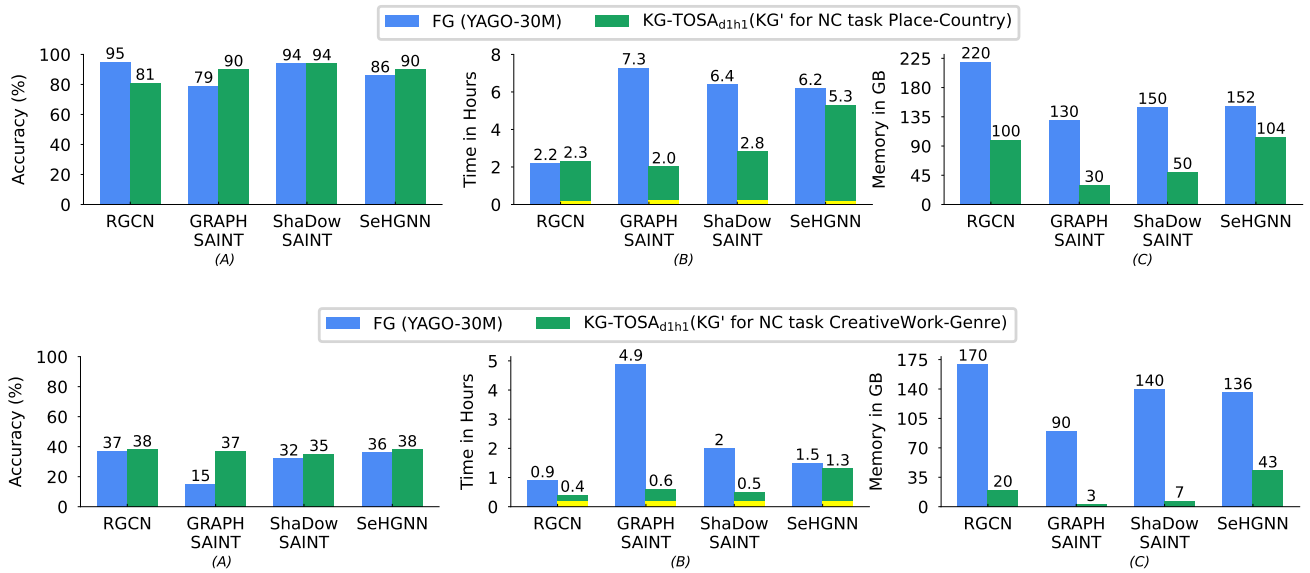

Figure 2: Performance of RGCN, Graph-SAINT, ShaDow-SAINT, and SeHGNN in the NC tasks. (A) Accuracy (higher is better), (B) Training-Time (lower is better), (C) Training-Memory (lower is better). The figures on top show the results for the place-country classification task on YAGO. The figures at the bottom show the results for the CreativeWork-Genere classification task. KG-TOSA enables all methods to reduce memory and time while improving accuracy , even with KG-TOSA's preprocessing time in yellow.

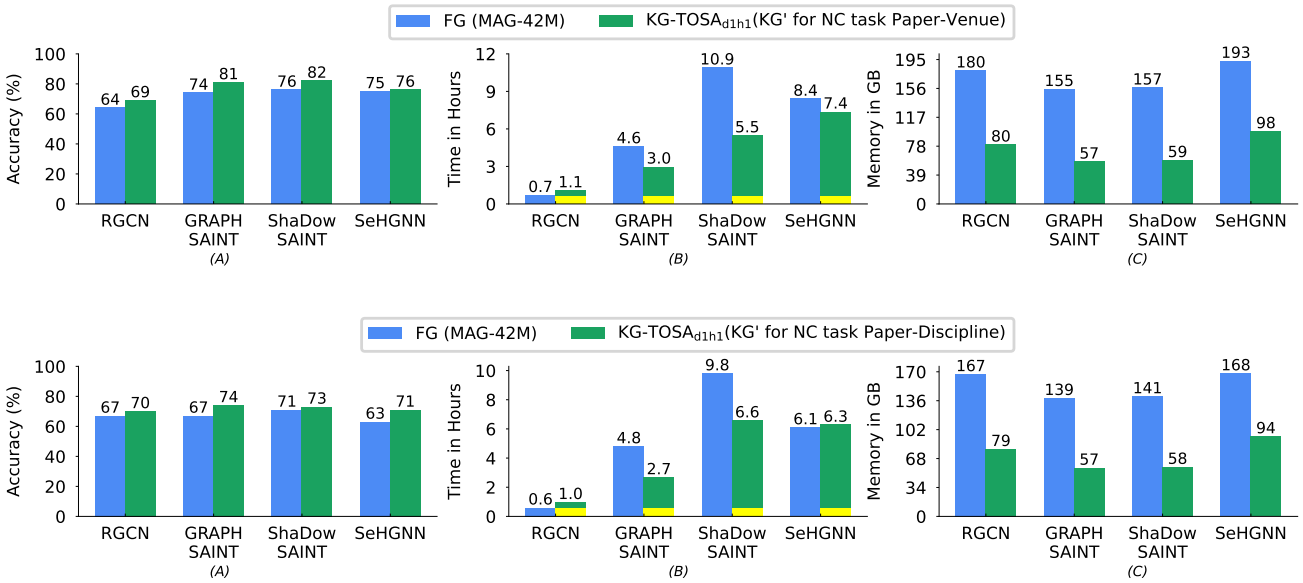

Figure 3: Performance of RGCN, Graph-SAINT, ShaDow-SAINT, and SeHGNN in the NC tasks. (A) Accuracy (higher is better), (B) Training-Time (lower is better), (C) Training-Memory (lower is better). The figures on top show the results for the paper-venue classification task on MAG. The figures at the bottom show the results for the Paper-Discipline classification task on MAG. KG-TOSA enables all methods to reduce memory, time and accuracy, even with KG-TOSA's preprocessing time in yellow

3 TRAINING CONVERGENCE RATE RESULTS

Figure 4 shows the convergence rate of RGCN (50 epochs) on the six node classification tasks.

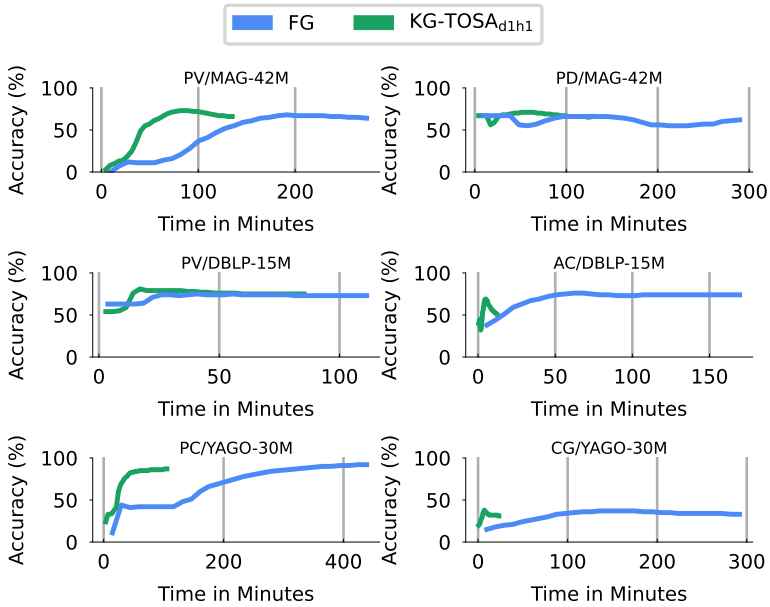

Figure 4: Convergence rate analysis. RGCN while training the six NC tasks using the full graph (FG) and KG' extracted by KG-TOSA. KG-TOSA enables the GNN method to generalize faster with comparable accuracy.

REFERENCES

[1] Marcel R. Ackermann. 2022. *dblp in RDF*. Retrieved June 07, 2022 from https://blog.dblp.org/2022/03/02/dblp-in-rdf/

[2] Mingyang Chen, Wen Zhang, Yushan Zhu, Hongting Zhou, Zonggang Yuan, Changliang Xu, and Huajun Chen. 2022. Meta-Knowledge Transfer for Inductive Knowledge Graph Embedding. In *Proceedings of the 45th International ACM SIGIR Conference (SIGIR '22)*. Association for Computing Machinery, 927–937. https://doi.org/10.1145/3477495.3531757

[3] Michael Färber. 2019. The Microsoft Academic Knowledge Graph: A Linked Data Source with 8 Billion Triples of Scholarly Data. In *The Semantic Web - ISWC, Proceedings, Part II (Lecture Notes in Computer Science)*, Vol. 11779. 113–129. https://doi.org/10.1007/978-3-030-30796-7_8

[4] Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. 2020. Open Graph Benchmark: Datasets for Machine Learning on Graphs. In *Advances in Neural Information Processing Systems 33:NeurIPS*.

[5] Thomas Pellissier Tanon, Gerhard Weikum, and Fabian M. Suchanek. 2020. YAGO 4: A Reason-able Knowledge Base. In *The Semantic Web - 17th International Conference, ESWC (Lecture Notes in Computer Science)*, Vol. 12123. Springer, 583–596. https://doi.org/10.1007/978-3-030-49461-2_34

[6] Xiaocheng Yang, Mingyu Yan, Shirui Pan, Xiaochun Ye, and Dongrui Fan. 2023. Simple and Efficient Heterogeneous Graph Neural Network. *AAAI* abs/2207.02547 (2023). https://doi.org/10.48550/arXiv.2207.02547

[7] Hanqing Zeng, Muhan Zhang, Yinglong Xia, and et.al. 2022. Decoupling the Depth and Scope of Graph Neural Networks. *CoRR* abs/2201.07858 (2022). https://arxiv.org/abs/2201.07858