Task-Oriented Graph Neural Network Training on Large Knowledge Graphs for Accurate and Efficient Modeling

(Supplementary Material)

PVLDB Reference Format:
. Task-Oriented Graph Neural Network Training on Large Knowledge Graphs for Accurate and Efficient Modeling. PVLDB, 14(1): XXX-XXX, 2020.
doi:XX.XX/XXX.XX

PVLDB Artifact Availability:
The source code, data, and/or other artifacts have been made available at URL_TO_YOUR_ARTIFACTS.

1 OPEN KG BENCHMARK DATASETS

Existing GNN datasets are not suitable for evaluating our problem as they are either homogeneous graphs datasets as in [3] or heterogeneous datasets extracted from real KGs as in [3, 4] with small sizes and a few numbers of node/edge types [4].

1.1 Node Classification Datasets:

MAG-42M Paper-Venue (PV): The dataset extracted from The MAG KG[2] to train a model that predicts the publication venue of a paper based on its publication history. The target nodes are papers and the prediction labels are their associated venues. The set of paper is the same set in OGBN-MAG [3] dataset while including all neighbours nodes for a depth of 3 hops. The prediction task is a single-class classification task. the number of 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.

MAG-42M Paper-Discipline (PD): The dataset extracted from The MAG KG[2] to train a model that predicts the discipline of a paper based on its authors and citations. The target nodes are papers and the prediction labels are their associated disciplines. The set of paper is the same set in OGBN-MAG [3] dataset while including all neighbours nodes for a depth of 3 hops. Due to a large number of distinct disciplines, we choose the top 50 frequent disciplines only while any paper that does not belong to the top 50 is excluded. The prediction task is a single-class classification task and hence only the first associated with the paper is considered. 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.

Table 1: Statistics of KG Datasets (n-type: node type, e-type: edge type. The number of nodes and edges is in millions. The number of node/edge types in tens to hundred.

KG-Dataset	#nodes	#edges	#n-type	#e-type
MAG-42M	42.4M	166M	58	62
DBLP-15M	15.6M	252M	42	48
YAGO-30M	30.7M	400M	104	98
YAGO3-10	123K	1.1M	23	37
OBGN-MAG	1.9MK	62M	4	4

DBLP-15M Paper-Venue (PV): The dataset extracted from The DBLP KG[1] to train a model that predicts the publication venue of a paper based on its publication history. The target nodes are papers that have type *(rec)* in KG and the prediction labels are their associated venues i.e. *(predicate type publishedIn)*. The blank nodes are excluded. The prediction task is a single-class classification task. Due to a large number of distinct venues, we choose the top 50 frequent venues only while any paper that does not belong to the top 50 is removed 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. The number of node types is 42 and the number of edge types is 48. the number of edges is 252M and the number of nodes is 15.6M.

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 nodes are authors that have type *(person)* in KG and the prediction labels are their associated venues i.e. *(predicate type Affaliation_Country)*. The blank nodes are excluded. The prediction task is a single-class classification task. Due to a large number of distinct countries, we choose the top 50 frequent countries only while any author affiliation that does not belong to the top 50 is excluded. Suppose an author has multiple affiliations countries, the current affiliation is chosen. The data is randomly split into ratios of 90%, 10%, and 10% for training validation and test respectively.

Yago4-30M KG (PlaceCountry (PC)): The dataset extracted from The YAGO4 KG[5] to train a model that predicts the country of a place based on its neighbour places. The target nodes are places that have type *(place)* in KG and the prediction labels are their associated countries i.e. *(predicate type locatedIn)*. The blank nodes are excluded. The prediction task is a single-class classification task. Due to a large number of distinct countries, we choose the top 200 frequent countries only while any author place that does not belong to the top 200 is excluded. The data is randomly split into ratios of 90%, 10%, and 10% for training, validation and testing respectively.

Yago-4 KG (Creative workGenre (CG)): The dataset extracted from The YAGO4 KG[5] to train a model that predicts the genre of

Proceedings of the VLDB Endowment, Vol. 14, No. 1 ISSN 2150-8097.
doi:XX.XX/XXX.XX

Table 2: The Open knowledge graph benchmark (OKGB) Tasks. The GNN task type is either node classification (NC) or link prediction (LP)

Task	Task-Type	Dataset	Target Node/Edge	Classes	Train-Test Splitting	Evauation Metric
Paper-Venue	NC	MAG-42M	Paper	Venues	Time	Accuracy(%)
Paper-Discipline	NC	MAG-42M	Paper	Disciplines	Time	Accuracy(%)
Paper-Venue	NC	DBLP-15M	Publication	Venues	Time	Accuracy(%)
Author Country	NC	BDLP-15M	Author	Coutries	Random	Accuracy(%)
CreativeWork-Genere	NC	YAGO-30M	CreativeWork	Geners	Time	Accuracy(%)
Place-Country	NC	YAGO-30M	Place	Countries	Random	Accuracy(%)
Connected Airports	LP	YAGO3-10	Connected To	-	Random	MRR-Hits@10
Author Affiliations	LP	DBLP15M	AffiliatiotedWith	-	Random	MRR-Hits@10

creative work based on its neighbour nodes. The target nodes are creative workpieces that have type *(genere)* in KG and the prediction labels are their associated genres i.e. *(predicate type genre)*. The blank nodes are excluded. The prediction task is a single-class classification task. Due to a large number of distinct genres, we choose the top 51 frequent genres only while any creative work that does not belong to the top 51 is excluded. The data is randomly split into ratios of 90%, 10%, and 10% for training, validation and testing respectively.

1.2 Link Prediction Benchmark datasets:

Connected Airports Link: The dataset extracted from The YAGO3 KG[5] to train a model that predicts the missing *isConnected* link between two nodes of type airport. The target nodes are airports that have *(isConnected)* link in between in the KG. The blank nodes are excluded. The data is randomly split into ratios of 90%, 10%, and 10% for training, validation and testing respectively. **DBLP Author PrimaryAffiliations:** The dataset extracted from The DBLP KG[1] to train a model that predicts the missing *Primary-Affiliations* link between two nodes of type author and research institute. The target nodes are the author and research institute that have *(PrimaryAffiliations)* link in between in the KG. The blank nodes are excluded. The data is randomly split into ratios of 90%, 10%, and 10% for training, validation and testing respectively.

2 NODE CLASSIFICATION RESULTS

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

3 TRAINING CONVERGENCE RATE RESULTS

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

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] 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

[3] 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.*

[4] Qingsong Lv, Ming Ding, and et. al. 2021. Are we really making much progress?: Revisiting, benchmarking and refining heterogeneous graph neural networks. In *KDD 21.* ACM, 1150–1160. https://doi.org/10.1145/3447548.3467350

[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

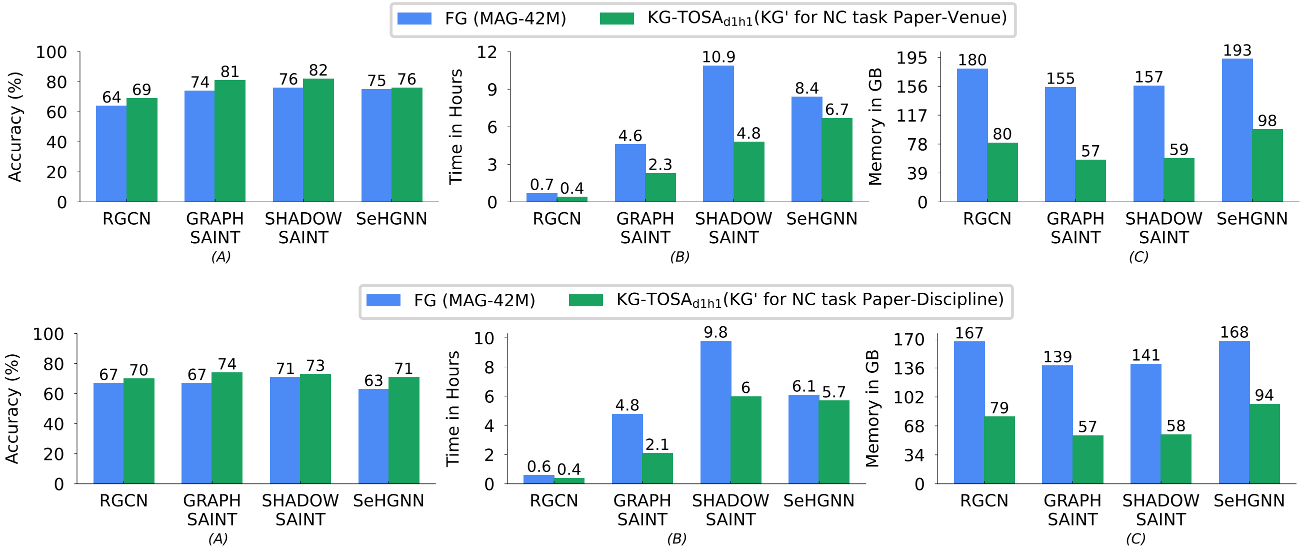

Figure 1: Performance of RGCN, Graph-SAINT, ShaDow-SAINT, and SeHGNN in the NC tasks. (A) Accuracy (higher is better), (B) Training-Time (lower is better), (C) Training-Memory (lower is better). The figures on top show the results for the paper-venue classification task on MAG. The figures at the bottom show the results for the Paper-Discipline classification task on MAG. I

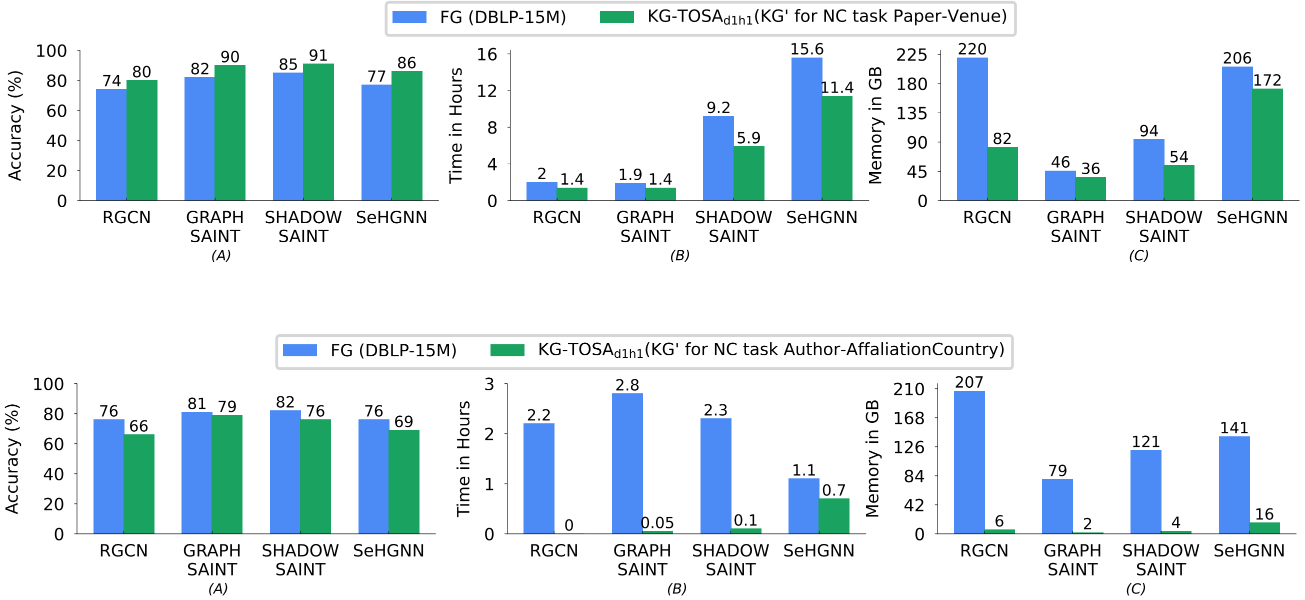

Figure 2: Performance of RGCN, Graph-SAINT, ShaDow-SAINT, and SeHGNN in the NC tasks. (A) Accuracy (higher is better), (B) Training-Time (lower is better), (C) Training-Memory (lower is better). The figures on top show the results for the paper-venue classification task on DBLP. The figures at the bottom 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.

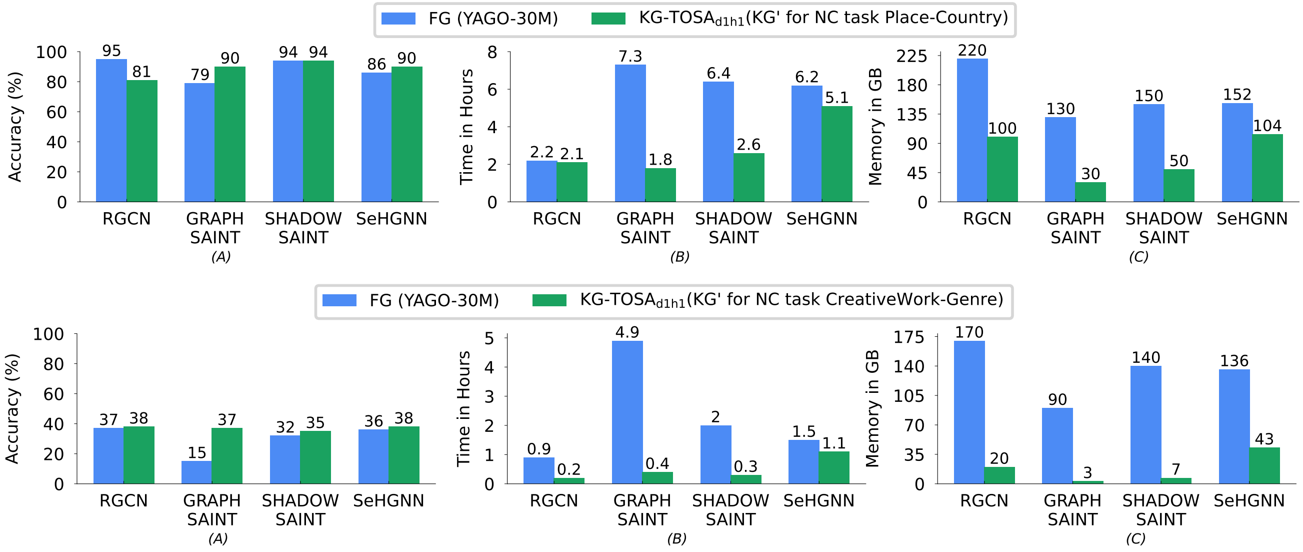

Figure 3: Performance of RGCN, Graph-SAINT, ShaDow-SAINT, and SeHGNN in the NC tasks. (A) Accuracy (higher is better), (B) Training-Time (lower is better), (C) Training-Memory (lower is better). The figures on top show the results for the place-country classification task on YAGO. The figures at the bottom show the results for the CreativeWork-Genere classification task. KG-TOSA enables all methods to reduce memory and time while improving accuracy in most methods.

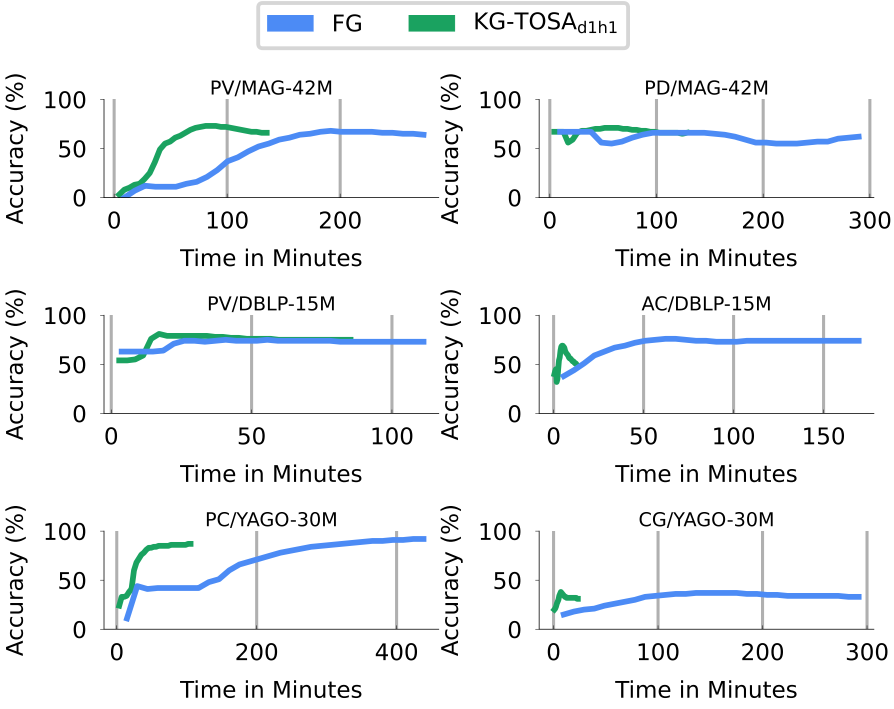

Figure 4: Convergence rate analysis. RGCN while training the six NC tasks using the full graph (FG) and KG' extracted by KG-TOSA. KG-TOSA enables the GNN method to generalize faster with comparable accuracy.