EXPERT: Public Benchmarks for Dynamic Heterogeneous Academic Graphs

Sameera Horawalavithana, Ellyn Ayton, Anastasiya Usenko, Shivam Sharma,
Jasmine Eshun, Robin Cosbey, Maria Glenski, and Svitlana Volkova
Pacific Northwest National Laboratory
firstname.lastname@pnnl.gov

ABSTRACT

Machine learning models that learn from dynamic graphs face nontrivial challenges in learning and inference as both nodes and edges change over time. The existing large-scale graph benchmark datasets that are widely used by the community primarily focus on homogeneous node and edge attributes and are static. In this work, we present a variety of large scale, dynamic heterogeneous academic graphs to test the effectiveness of models developed for multi-step graph forecasting tasks. Our novel datasets cover both context and content information extracted from scientific publications across two communities – Artificial Intelligence (AI) and Nuclear Nonproliferation (NN). In addition, we propose a systematic approach to improve the existing evaluation procedures used in the graph forecasting models.

ACM Reference Format:
Sameera Horawalavithana, Ellyn Ayton, Anastasiya Usenko, Shivam Sharma,, Jasmine Eshun, Robin Cosbey, Maria Glenski, and Svitlana Volkova . 2022. EXPERT: Public Benchmarks for Dynamic Heterogeneous Academic Graphs. In *Proceedings of Workshop on Graph Learning Benchmarks, Web Conference (GLB '22).* ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

The adoption of several benchmark datasets by the graph-based learning community has spurred many research challenges and opportunities in large-scale graph modeling [7] and out-of-distribution generalization [6, 10]. These benchmarks are useful across inference tasks and domains to understand the limitations of existing Graph Neural Network (GNN) models or to develop new GNN models for state-of-the-art graph-based learning tasks, *e.g.,* link prediction. For example, the top-ranked solutions of the 2021 KDD cup developed large and deep GNN models to predict primary subject areas of Arxiv papers released in the Open Graph Benchmark (OGB) [6]. Another benchmark study revealed the limitations of existing GNN methods in non-homophilous graphs [10].

Although most of the existing large-scale graph benchmark datasets focus on static homogeneous graphs [6, 10], many real-world problems involve dynamic graphs where heterogeneous nodes and edges change over time. The evolving nature of dynamic graphs requires handling new, previously unseen nodes as well as capturing temporal patterns. Several promising graph-based learning frameworks have been developed for dynamic heterogeneous graphs [9, 11, 16]. The proposed approaches learn time-aware node embeddings to represent the evolving topological structures. However, these methods are yet to be tested on standard large-scale dynamic graph benchmark datasets. Scholarly publications present an avenue for studying dynamic heterogeneous graphs. Large-scale dynamic heterogeneous academic graphs capture scientific knowledge development and collaboration patterns across disciplines, *e.g.,* artificial intelligence, natural language processing, and can provide new insights on how careers evolve, how collaborations drive scientific discovery, and how scientific progress emerges [14].

There are two contributions of this work: (1) the public release seven novel dynamic, heterogeneous academic graph benchmark datasets for two research communities (artificial intelligence and nuclear nonproliferation); and (2) standardized evaluation procedures for forecasting on dynamic, heterogeneous context graphs. We investigate and draw novel insights about performance evaluation for graph forecasting tasks using a systematic approach analyzing the complexity of both transductive and inductive predictions.

2 RELATED WORK

In this section, we outline related work on academic graph benchmark datasets. Existing graph benchmarks primarily focus on either node classification or link prediction tasks. Cora [5] and CiteSeer [5] datasets are widely used [18] for node classification, where the task is to predict the missing subject area of a paper represented as a node in the graph. However, these datasets are very small ($< 4K$ nodes) which makes them unsuitable for evaluation of graph-based ML models; moreover, there have been several issues reported about data quality [3, 7]. Open Graph Benchmark (OGB) [7] released three academic graph benchmark datasets (*e.g.,* ogbn-arxiv, ogbn-papers100M, ogbn-mag) for node classification tasks. OGB datasets are extracted from the Microsoft Academic Graph (MAG) [15]. Several recent works show the usefulness of the OGB datasets for large-scale graph learning [6]. In link prediction tasks, evaluation is focused on predicting missing edges. DBLP [17] and ogbl-citation2 [7] are two commonly used benchmarks for link prediction in static citation networks. Similarly, HEP-PH [4] is a benchmark for link prediction in a citation network, but considers a dynamic graph setting. However, these datasets are not suited for evaluation on edge forecasting tasks on *heterogeneous* graphs.

As shown in Table 1, most existing large-scale graph benchmarks focus on static, homogeneous graphs and are mainly co-citation and

GLB '22, April 25–29, 2022, Virtual

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

collaboration networks. The most similar to our work is the *ogbn-mag* benchmark released as a part of the Open Graph Benchmark (OGB) [7]. This heterogeneous graph dataset includes several types of nodes (authors, institutions, papers and topics) with multiple edge types – affiliations, authorships, citations and paper-topic relationships. The *ogbn-mag* prediction task is to predict the missing venue (conference or journal). While this dataset is useful for the development of heterogeneous graph-based models, it does not take into account the timestamped edges. Unlike any previous work, we present a novel dynamic heterogeneous academic graphs dataset representing seven data sources across two domains.

Table 1: Existing benchmarks for academic graph datasets, including whether graphs are dynamic (D). We use OGB statistics reported in the original paper [7].

Benchmark	# Nodes	# Edges	D	Edge Types	Pred. Task
Cora [5]	2.7K	5.4K	–	citation	subject areas
CiteSeer [5]	3.3K	4.7K	–	citation	subject areas
ogbn-arxiv [7]	0.2M	1.1M	–	citation	subject areas
ogbn-papers100M [7]	111M	1.6B	–	citation	subject areas
ogbl-collab [7]	0.2M	1.3M	–	collaboration	collaborations
ogbl-citation2 [7]	2.9M	30.6M	–	citation	citations
DBLP [17]	0.3M	1.1M	–	citation	citations
ogbn-mag [7]	1.9M	21M	–	affiliations, authorship, citations, topics	venue
HEP-PH [4]	34K	0.4M	✓	citation	citations
Our Work*	3.5M	34M	✓	collaboration, partnership, expertise	collaboration, partnership, expertise

3 GRAPH DATASET CONSTRUCTION

We make available[1] dynamic heterogeneous academic graphs for seven sources across two research communities — artificial intelligence (AI) and nuclear nonproliferation. Graphs are split temporally into train, validation, and test sets. We reserve the last year(s) of data as the test set, depending on venue scale. The year prior to test is used for validation and remaining years for training. A summary of node and edge types are presented for each data split in Table 2.

Data Collection. AI papers were collected from the proceedings of four top-tier AI conferences: Association for Computational Linguistics (ACL), the International Conference on Machine Learning (ICML), the International Conference on Learning Representations (ICLR) and the Conference on Neural Information Processing Systems (NeurIPS). We downloaded publication PDFs, which we parsed using GROBID [1] and CERMINE [12] to extract content as well as metadata (author affiliations, country affiliations, etc.). To obtain focused benchmarks, we filtered publications using a set of keywords curated by subject matter experts (SMEs)[2].

[1]All data is available from our GitHub and hosted by the Berkeley Data Cloud (https://bdc.lbl.gov) under the *Global Expertise Forecasting* project

[2]AI keywords: fair, ethic, translation model, machine translation, dialog, genetic algorithm, explanation, transfer learning, clustering, adversarial, nlg, sentiment, causal, reinforcement learning, transparent, summarization, question-answer, interpretability, language model, interpretable

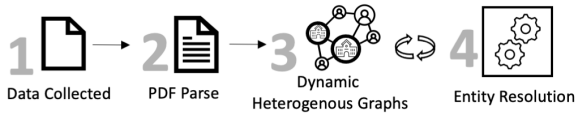

Figure 1: Academic graph data pre-processing pipeline.

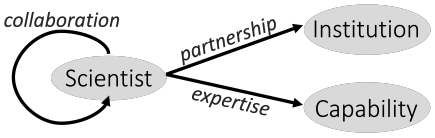

Figure 2: Schema of our dynamic academic graphs.

Nuclear publication records and abstracts were collected from three sources: the Office of Scientific and Technical Information (OSTI), Web of Science (WoS), and the SCOPUS database using a set of nuclear keywords[3] compiled with domain resources[4] and manual specification from SMEs. False positives were removed using topic modeling [2] and SME verification of non-nuclear topics.

Processing and Graph Construction. Using an in-house developed processing pipeline outlined in Figure 1, we processed publication meta-data (aka context) and into dynamic heterogeneous academic graphs (schema in Figure 2). For every dataset, we performed entity resolution over all scientists and institutions. Given the many ways author and institution names can be represented (*e.g.,* Jane Doe and J. Doe), we created unique nodes for authors and affiliations in each paper and manually combined nodes if two entities were determined to be the same.

When resolving nodes, we consider text similarity, the edit distance between the two names, and graph similarity of nodes' ego networks. For example, the two scientists, Jane Doe and J. Doe, have a high text similarity since their names are almost identical. These scientists are merged if they also have a high graph similarity score, *i.e.,* similar coauthors, capabilities and institutions. Leveraging both metrics prevents us from incorrectly merging scientists, *e.g.,* Jane Doe and John Doe. We iteratively apply this process to the graphs until only node pairs with low similarity scores are returned. Nuclear datasets with >500K nodes were partially manually resolved due to size. In this case, we applied an automatic resolver on nodes with similarity scores above a threshold heuristically determined for each dataset. Processing code is available on our GitHub[5].

4 TASK FORMULATION

We introduce several questions of interest that can be answered by studying our proposed dynamic heterogeneous academic graphs. We group them into three categories: collaboration, partnership, and capability evolution.

Collaborations. Scientific publications are authored by individuals or teams of scientists. Previous work [14] has shown that team-authored publications are more popular in terms of citations

[3]Nuclear keywords from: github.com/pnnl/expert/blob/master/expert/queries.py

[4]IAEA Safety Glossary: www-pub.iaea.org/MTCD/Publications/pdf/PUB1830_web.pdf

[5]Data processing code: https://github.com/pnnl/EXPERT/tree/master/examples

Table 2: Characteristics of the datasets used in the experiments during training and evaluation.

	Graph	*Data Split*	*Time Range*	*# Nodes*			*# Edges*		
				#Scientists	#Institutions	#Capabilities	#Collaborations	#Partnerships	#Expertise
AI Domain	**ACL**	Training	1965-2018	38,436	7,979	20	237,174	207,096	10,430
		Validation	2019	9,691	1,966	19	45,231	31,989	1,941
		Testing	2020-2021	8,386	1,848	20	34,237	28,909	1,103
	ICML	Training	2009-2019	9,229	493	20	5,905	20,672	12,599
		Validation	2020	3,980	281	19	2,297	8,032	4,581
		Testing	2021	4,564	301	18	2,552	10,088	5,199
	ICLR	Training	2016-2019	5,272	490	19	3,900	11,934	5,797
		Validation	2020	2,639	276	20	1,925	6,559	2,916
		Testing	2021	3,407	284	20	2,498	9,006	3,881
	NeurIPS	Training	1987-2018	22,150	1,441	20	12,727	31,059	22,619
		Validation	2019	4,572	454	19	2,934	9,256	4,810
		Testing	2020	7,968	572	20	4,918	16,962	8,570
Nuclear Domain	**WoS**	Training	2015-2018	1,309,530	108,562	61	4,075,741	10,903,275	156,476
		Validation	2019	449,094	48,556	47	1,280,842	3,026,751	37,868
		Testing	2020	311,206	34,021	47	867,578	2,076,699	23,186
	SCOPUS	Training	2015-2018	222,051	88,758	43	1,182,054	2,399,450	460,579
		Validation	2019	91,614	34,385	42	387,267	748,045	143,513
		Testing	2020-2021	62,665	26,061	40	258,153	499,326	91,117
	OSTI	Training	2015-2018	247,724	17,303	43	176,395	1,203,955	520,978
		Validation	2019	25,008	2,584	39	38,014	337,073	46,022
		Testing	2020	47,222	4,994	41	119,676	1,363,581	119,676

than single-authored publications. Questions focus on underlying patterns of collaboration: Are there persistent groups of scientists who collaborate repeatedly? Do veteran scientists collaborate with early career or veteran scientists? Do collaborations occur within tightly connected groups of scientists?

Partnerships. Scientists may collaborate with other scientists from the same institution or across multiple. What drives such multi-institutional partnerships? Are researchers at elite universities more likely to collaborate with scientists at other elite universities? To what extent do scientists collaborate internationally? Are papers authored by international groups of scientists more likely to be published in high-impact journals?

Capability Evolution. Teams of scientists produce diverse but specialized capabilities in comparison to what any individual collaborator could produce [14]. Are there differences in the topics that scientists tend to tackle? Are scientists going to adopt the most recent and emerging capabilities? Which scientists will disrupt science by suggesting new tasks and opening up novel opportunities? How often do scientists generate more theoretical innovations in contrast to empirical analyses?

4.1 Multi-step Link Prediction Task

We consider dynamic heterogeneous academic graphs (G) consisting of scientists, institutions, and capabilities as nodes (N). A pair of nodes is connected at a timestamp (t), by a directed edge which is denoted by a quadruple (N_s, r, N_o, t). Edges are of multiple types (r) such as collaboration (*scientist-to-scientist*), partnership (*scientist-to-institution*) and proficiency (*scientist-to-capability*). An ordered sequence of quadruples represents the dynamic heterogeneous graph. In contrast to predicting missing edges in a static graph (*interpolation*), we need to predict the future edges in a dynamic graph (*extrapolation*). As these edges occur over multiple time stamps in the future, we treat the prediction task as a multi-step inference task (see Definition 1). Thus, we need to develop methods that can extrapolate the graph structure over future timestamps [9]. In our use case, such predictions are useful to forecast emerging science trends in terms of global expertise and capability development.

Definition 1. *Given a graph (G_T) that represents the ordered sequence of quadruples until time T, the task is to forecast the graph ($G_{T:T+m}$) over multiple future time steps (m). G_T can be represented as discrete-time dynamic graphs (e.g., sequences of static graph snapshots) and continuous-time dynamic graphs (e.g., timed lists of edges).*

4.2 Task Complexity

In this section, we discuss temporal link prediction as transductive and inductive tasks as illustrated in the taxonomy in Figure 3. Our objective is to understand the complexity of different setups.

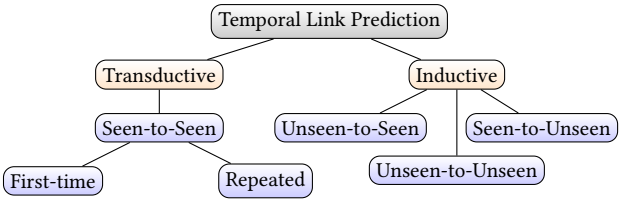

Figure 3: Forecasting Task Taxonomy.

We group the edges in the test data into multiple categories within transductive (test sets only include edges between nodes *seen* in training) and inductive (there is at least one "unseen" node in the test edges that was not present in training data) settings.In the transductive setting, edges capture incumbent scientists who

publish in the same venue repeatedly and we group interactions between scientists into "First-time" and "Repeated" categories. In the inductive setting, there are three groups of interactions: "Unseen-to-Seen", "Unseen-to-Unseen", and "Seen-to-Unseen". For example, a graduate student ("unseen") can publish her first paper in the ACL community with her mentor ("seen"), or a group of scientists may publish in the ACL community for the first time ("Unseen-to-Unseen"). We distinguish "Unseen-to-Seen" and "Seen-to-Unseen" interactions between a "seen" and an "unseen" node input to the model as the directionality makes these two variants different in nature and predictive context; *e.g.,* a new graduate student may interact only with their mentor.

We use these edge groups to understand the complexity of temporal link prediction task in both settings. We define the task complexity in the transductive setting as the proportion of "First-time" interactions with respect to the "Repeated" interactions. We take the proportion of "Seen-to-Unseen", "Unseen-to-Seen" and "Unseen-to-Unseen" interactions with respect to other edges to define the complexity of the inductive prediction task. The higher the proportion, the higher the complexity of both transductive and inductive tasks. As shown in Figure 4, we notice that the complexity of the transductive task increases over different train/test splits over time (for both ACL and WoS graphs, there are emerging interactions between incumbent scientists) while inductive task complexity remains comparable. While there are many new authors who publish over time in the ACL data, they are relatively low in the WoS data.

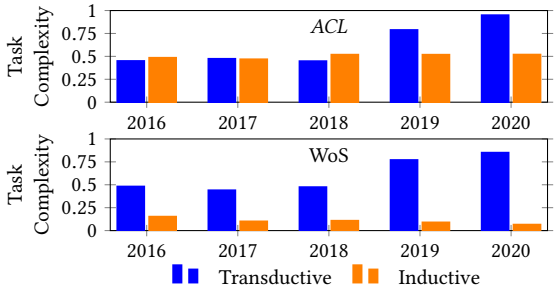

Figure 4: Transductive and inductive task complexity.

4.3 Design of Evaluation

In this section, we propose a systematic approach to improve the current evaluation procedures. Previous work considers multi-step graph forecasting as a variant of the temporal link prediction task where interactions are either ranked or classified [11, 13]. In a ranking solution, the model ranks the potential nodes (N_o) that would interact with a given node (N_s), relationship (r), and timestamp (t). Metrics typically include mean rank (MR), mean reciprocal rank (MRR), and the percentage of examples with the true target entity in top K candidates (known as Hits@K). There are a few limitations in this approach. First, we can only rank target nodes that are known (*i.e.,* seen in training data). Second, given a training graph with N nodes, we need to make N^2 inferences at most; This is intractable for very large graphs with millions of nodes. Most recent works attempt to rank only a subset of target nodes ($M << N$) to reduce the number of inferences ($M \times N <<< N^2$). However, performance heavily depends on the chosen subset of target nodes.

When treated as a classification task, models predict the existence of an edge (N_s, r, N_o) at future time t. Metrics include precision, recall, and AUC (area under the receiver operating characteristic curve). This approach would be suitable for both transductive and inductive prediction tasks as its objective is to distinguish interactions from the non-interactions. A common approach is to construct a set of negative examples equal to the number of positive examples in test sets, and thus performance can be heavily influenced by those chosen negative examples.

Evaluation Recommendations. First, we propose to interpret the transductive and inductive task performance in context of our task complexity measure. This indicates whether a model is capable of capturing the distribution shifts observed on the training and testing splits. Distribution shifts may create very different patterns of collaboration, partnership, and capability development in the testing period than in the training period. However, detecting such a distribution shift or determining what causes a shift is hard [8].

Second, we recommend to report link prediction performance across multiple edge types. For example, a model may perform well predicting *collaboration* edges, but may not perform comparably predicting the *scientist to capability* edges. This nuanced evaluation feedback can be used to target model improvement to boost overall performance or generalizability. Link prediction tasks should also be evaluated separately for the transductive and inductive settings, as discussed in Section 4.2. We notice that many recent works filter the test edges that co-appear in the train, valid, or test sets in the evaluation. While the intention may be to focus on performance for unseen relationships, this is operationally irrelevant. For example, one may be interested in predicting whether a group of scientists would repeatedly publish on the same conference. These groups are persistent as they would appear on both train and test splits.

Third, nodes that do not receive updates regularly over time need to be explicitly accounted for. There may be scientists who do not publish regularly in the same venue. In this case, nodes may include gaps of activity in training data. Predicting for these inactive/inconsistent scientists would be more challenging due to thee unusual activity flow compared to active scientists. We need to design systems that would address this staleness problem.

Finally, our new datasets can support link prediction across multiple future time steps to enable evaluation of performance differences over time and with varying prediction windows. There may be a drop in the performance with increasing time steps, or increasing gaps between train and test periods.

5 SUMMARY AND CONCLUSIONS

In this paper, we release seven dynamic heterogeneous academic graphs benchmark datasets to understand how scientific collaboration, partnership and authorship evolve in AI and nuclear nonproliferation communities. These graph datasets consist of 3.5M nodes and 34M timestamped edges in total and we show the complexity of transductive and inductive tasks through a systematic approach. We hope our contributions will help researchers to build and evaluate new graph models, or understand the limitations of existing graph models in dynamic heterogeneous graph forecasting tasks.

ACKNOWLEDGMENTS

This material is based on work funded by the United States Department of Energy (DOE) National Nuclear Security Administration (NNSA) Office of Defense Nuclear Nonproliferation Research and Development (DNN R&D) Next-Generation AI research portfolio and Pacific Northwest National Laboratory, which is operated by Battelle Memorial Institute for the U.S. Department of Energy under contract DE-AC05-76RLO1830. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the United States Government or any agency thereof. Authors thank Joonseok Kim and Sannisth Soni for their help with preparing the datasets.

REFERENCES

[1] 2008–2021. GROBID. https://github.com/kermitt2/grobid.
[2] Dimo Angelov. 2020. Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470* (2020).
[3] Vijay Prakash Dwivedi, Chaitanya K Joshi, Thomas Laurent, Yoshua Bengio, and Xavier Bresson. 2020. Benchmarking graph neural networks. *arXiv preprint arXiv:2003.00982* (2020).
[4] Johannes Gehrke, Paul Ginsparg, and Jon Kleinberg. 2003. Overview of the 2003 KDD Cup. *Acm Sigkdd Explorations Newsletter* 5, 2 (2003), 149–151.
[5] Lise Getoor. 2005. Link-based classification. In *Advanced methods for knowledge discovery from complex data*. Springer, 189–207.
[6] Weihua Hu, Matthias Fey, Hongyu Ren, Maho Nakata, Yuxiao Dong, and Jure Leskovec. 2021. Ogb-lsc: A large-scale challenge for machine learning on graphs. *arXiv preprint arXiv:2103.09430* (2021).
[7] 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. *arXiv preprint arXiv:2005.00687* (2020).
[8] C. Huyen. 2022. *Designing Machine Learning Systems: An Iterative Process for Production-Ready Applications*. O'Reilly Media, Incorporated. https://books.google.com/books?id=70KmzgEACAAJ
[9] Woojeong Jin, Meng Qu, Xisen Jin, and Xiang Ren. 2019. Recurrent event network: Autoregressive structure inference over temporal knowledge graphs. *arXiv preprint arXiv:1904.05530* (2019).
[10] Derek Lim, Xiuyu Li, Felix Hohne, and Ser-Nam Lim. 2021. New benchmarks for learning on non-homophilous graphs. *arXiv preprint arXiv:2104.01404* (2021).
[11] Emanuele Rossi, Ben Chamberlain, Fabrizio Frasca, Davide Eynard, Federico Monti, and Michael Bronstein. 2020. Temporal graph networks for deep learning on dynamic graphs. *arXiv preprint arXiv:2006.10637* (2020).
[12] Dominika Tkaczyk, Paweł Szostek, Mateusz Fedoryszak, Piotr Jan Dendek, and Łukasz Bolikowski. 2015. CERMINE: automatic extraction of structured metadata from scientific literature. *International Journal on Document Analysis and Recognition (IJDAR)* 18, 4 (2015), 317–335.
[13] Francesco Visin, Kyle Kastner, Kyunghyun Cho, Matteo Matteucci, Aaron Courville, and Yoshua Bengio. 2015. Renet: A recurrent neural network based alternative to convolutional networks. *arXiv preprint arXiv:1505.00393* (2015).
[14] Dashun Wang and Albert-László Barabási. 2021. *The science of science*. Cambridge University Press.
[15] Kuansan Wang, Zhihong Shen, Chiyuan Huang, Chieh-Han Wu, Yuxiao Dong, and Anshul Kanakia. 2020. Microsoft academic graph: When experts are not enough. *Quantitative Science Studies* 1, 1 (2020), 396–413.
[16] Da Xu, Chuanwei Ruan, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. 2020. Inductive representation learning on temporal graphs. *arXiv preprint arXiv:2002.07962* (2020).
[17] Jaewon Yang and Jure Leskovec. 2015. Defining and evaluating network communities based on ground-truth. *Knowledge and Information Systems* 42, 1 (2015), 181–213.
[18] Zhilin Yang, William Cohen, and Ruslan Salakhudinov. 2016. Revisiting semi-supervised learning with graph embeddings. In *International conference on machine learning*. PMLR, 40–48.