Heterogeneous Graph Representation for Dataset Link Prediction on Dynamic and Sparse Scholarly Graphs

Ornella Irrera[1](✉), Matteo Lissandrini[2], Daniele Dell'Aglio[3], and Gianmaria Silvello[1]

[1] Department of Information Engineering, University of Padova, Padova, Italy
{ornella.irrera,gianmaria.silvello}@unipd.it

[2] Department of Foreign Languages and Literatures, University of Verona, Verona, Italy
matteo.lissandrini@univr.it

[3] Department of Computer Science, Aalborg University, Aalborg, Denmark
dade@cs.aau.dk

Abstract. Scientific data are crucial for conducting and validating research, yet they are often undervalued and poorly integrated within the broader scientific ecosystem. This issue is reflected in the typically inadequate documentation of datasets and their weak connections to other research outputs in Scholarly Knowledge Graphs (SKGs).

Real-world SKGs present several challenges, including their large scale, heterogeneity (with nodes such as authors, venues, papers, and datasets), sparsity, and incompleteness (e.g., partial or missing descriptive nodes' metadata). SKGs are also dynamic, constantly evolving as new entities are introduced.

This paper presents a novel method for heterogeneous graph representation designed to improve publication-dataset link prediction – crucial for enhancing data discoverability and reuse. Our approach effectively addresses the challenges outlined, ensuring suitability for inductive settings. Extensive evaluations demonstrate that our method outperforms state-of-the-art techniques, showcasing its robustness and effectiveness in a wide range of scenarios. This makes it a viable solution for real-world applications, where it can contribute to improving search and access to scientific data within SKGs.

Keywords: Dataset link prediction · Scholarly knowledge graphs · Representation learning

1 Introduction

Sharing and reusing scientific data – encompassing observations, experimental results, and research-derived information – is now a key part of the scientific progress [4,26,36]. Scientific data allow for validating hypotheses, powering experiments, training AI algorithms, and accelerating research advancements. However, the historical absence of widely accepted practices for sharing,

W.-T. Balke et al. (Eds.): TPDL 2025, LNCS 16097, pp. 452–469, 2026.
https://doi.org/10.1007/978-3-032-05409-8_26

describing, and citing data has led to poorly documented datasets that are hard to find and only loosely connected to related research. As a result, many scientific datasets remain underappreciated and underutilized by the scientific community [5,6,26]. While proper data citation is crucial for giving credit to data curators, robust *publication-dataset linking* methods are essential for locating relevant datasets and connecting them to publications [9], enhancing the visibility of both datasets and their associated authors, improving reusability, and amplifying their overall impact on scientific research.

In this context, Scholarly Knowledge Graphs (SKGs), heterogeneous attributed graphs interconnecting research outcomes, are crucial for efficiently collecting, organizing, and exploring the vast amount of scholarly data – such as publications, authors, and keywords – and their interconnections. Several SKGs, such as the Microsoft Academic Knowledge Graph (MAKG) [32], Open Research Knowledge Graph (ORKG) [20], and the OpenAIRE Graph (OAG) [23], represent datasets along with other entities. These graphs are characterized by their *dynamic nature* (with new publications and datasets continually being added), *incompleteness* (with publications typically having rather complete metadata while datasets often lack adequate descriptions), and *sparsity* (loosely connected publications and datasets). Isolated or loosely connected datasets in SKGs, combined with incomplete and low-quality descriptive metadata, make these datasets difficult to discover and reuse effectively [18]. To address these issues, there is an urgent need for advanced automatic methods to infer missing links between publications and datasets, thereby enhancing SKG connectivity and dataset discoverability. These methods must be *robust* to manage heterogeneity, noise and sparsity, *adaptable* to leverage graph topology when text is lacking and *versatile* to eliminate the need for costly retraining whenever new data is added. Establishing links between publications and datasets is essential for building dense, interconnected research networks that support a variety of downstream tasks. No method has been developed explicitly for connecting publications with datasets. The most advanced approaches in the scholarly domain leverage Graph Representation Learning (GRL) [13,24,37], which transform graph elements into continuous vector spaces to capture structural and semantic properties. While not explicitly designed for publication-dataset link prediction, some GRL methods can be adapted for this purpose. These methods include both *homogeneous approaches* [16,21,30], which assume uniform node and edge types, and *heterogeneous approaches* [7,8,13,24,35,37], which are tailored for graphs with diverse type of nodes and relationships.

In this work, we propose the Scholarly Attention Network (SAN) method and evaluate its performance on dataset link prediction. SAN leverages heterogeneous graph representation and is designed to operate in real-world settings, accommodating heterogeneous, sparse, and noisy SKGs. Additionally, SAN works in both transductive and inductive scenarios – a capability tested for the first time here – and is effective whether full, or incomplete metadata are available. For evaluation, we rely on subsets of OAG, the backbone of the European Open Sci-

ence Cloud (EOSC, https://eosc.eu/), which is open-access, frequently updated, and comprizes scientific datasets.
We address three research questions:

RQ1 (Robustness). How well can dataset link prediction methods cope with the noise and sparsity commonly found in real-world SKGs?
RQ2 (Adaptability). Is it possible to strengthen the role of graph structure when textual metadata is limited?
RQ3 (Versatility). Can the models perform reliably in both transductive and inductive learning scenarios?

The main **contributions** of this work are as follows:

- Introduction of SAN, a novel heterogeneous graph representation method that demonstrates effectiveness and stability across transductive, inductive, and semi-inductive scenarios.
- A comprehensive evaluation across environments varying from metadata-rich to metadata-free, showing that SAN performs well in situations where text is scarce by effectively utilizing graph topology.
- A comparative analysis against existing graph-based baselines, revealing consistent improvements across all evaluated settings.

Outline. In Sect. 2 we describe current state-of-the-art GRL methods; in Sect. 3 we formalize the dataset link prediction task; in Sect. 4 we describe the SAN method; in Sects. 5 and 6 we describe the experimental setup and discuss the results. In Sect. 7, we report an ablation study. Finally, in Sect. 8, we draw some conclusions.

2 Related Works

2.1 Scholarly Knowledge Graphs (SKGs)

SKGs are structured representations of scientific outputs, interconnecting entities such as publications, authors, and datasets. Despite the growing recognition of the value of datasets, several scholarly graphs, like AMiner [27] and DBLP [22], primarily focus on mapping connections between publications.

In contrast, more recent SKGs have started incorporating datasets, such as the MAKG [32], which includes 200 million nodes and over 10 billion relationships; the ORKG [20], which contains about 5 million nodes and 50 million relationships; the DSKG [11], linking 2,000 datasets to 635,000 publications (and other Linked Data sources like MAKG, ORCID, and Wikidata); and the OAG [23], the largest with 227 million nodes and 15 billion relationships, offering open access to more than 60 million datasets. Recently, the MES graph [19], a subgraph of OAG, was published. It is semi-automatically curated and serves as a reliable ground truth dataset for testing graph-based machine learning algorithms.

On the other hand, we found that several state-of-the-art methods have been evaluated on resources that were excluded from our study due to various limitations – such as a high degree of homogeneity, task-specific design, reliance solely on metadata for prediction and recommendation, or a lack of representation of real-world SKGs. Specifically: (i) LinearSVM_Dataset is a bipartite graph with only $1,691$ dataset titles (from DSKG [11]) and $88K$ abstracts from MAKG, offering limited structural and relational diversity; (ii) The DataFinder Dataset [31] is not a proper SKG, as it lacks critical node types such as authors, venues, and organizations, consisting solely of 17K textual queries paired with 7K relevant datasets; (iii) The Delve [1] subsets, include datasets linked to an average of two publications each – a connectivity level that does not reflect the higher sparsity observed in real-world SKGs, where most datasets are linked to only a single publication.

2.2 Publication-Dataset Link Prediction

To date, no existing method specifically addresses the problem of interlinking publications and datasets. A straightforward approach could involve generating separate embeddings for publications and datasets, and then combining them to assess their mutual relevance and establish meaningful connections between them. Several methods exist to generate the embeddings. Techniques like Node2Vec [15] and Metapath2Vec [10] capture structural features through graph exploration and generate node embeddings. Translational models like TransE [3] represent relationships as vector translations, making them effective for capturing semantic similarity between entities. ComplEx [28] extends this by handling asymmetric relations using complex-valued embeddings. Other methods instead rely on GRL.

Graph Representation Learning (GRL). Representation learning in graphs aims at converting nodes, edges, or entire graphs into continuous vector spaces in a way that the resulting representations effectively capture the graph's structural, semantic, and relational information. These representations are used to perform various tasks, including node classification, link prediction, and recommendation.

Some approaches are designed for homogeneous graphs. Graph Convolutional Networks (GCNs) [21] focuses on learning node embeddings by aggregating and transforming information from neighboring nodes with graph convolutions. GraphSAGE [16] samples a fixed set of representative neighbors from each node and aggregates neighbors' representations based on mean, LSTM or pooling techniques. GAT [30] employs a similar method and proved to be effective on inductive settings by learning nodes representations via attention mechanism.

Other GRL techniques focus on heterogeneous graphs [34], which is more challenging due to the multiple types of nodes and edges, and varying feature sets across nodes. HetGNN [37] addresses these challenges by sampling neighbors for each type from a node's neighborhood and aggregating them using a bi-LSTM. RGCN [8] builds on GCNs by incorporating edge and node types. Other methods use attention mechanisms: GATNE [7] handles multiplex graphs by integrating edge attributes, while HGT [17] employs a transformer-based and type-specific attention mechanism to manage diverse node and edge types. HAN [35] and MAGNN [13] leverage *metapaths*, sequences of node and edge types that define relational patterns, with HAN aggregating information from different metapaths and MAGNN performing both intra- and inter-metapath aggregation. HiNormer [24] uses two encoders, a local structure encoder and a heterogeneous relation encoder.

HAN, HGT, HetGNN, MAGNN, and HiNormer have been tested on various subsets of AMiner and DBLP, but they considered only heavily cleaned subgraphs without dataset nodes. HetGNN and HGT were evaluated for link prediction (HetGNN for author-author and HGT for paper-author links).

3 Task Definition and Challenges

This section provides a formal definition of SKG, heterogeneous GRL, and the publication-dataset link prediction task with its open challenges.

Definition 1. *A Scholarly Knowledge Graph (SKG) is a tuple*

$$\mathcal{G}{:}(\mathcal{V}, \mathcal{E}, \mathcal{A}, \mathcal{R}, \alpha, \beta)$$

where $\mathcal{V}, \mathcal{E}, \mathcal{A}, \mathcal{R}$ *are the sets of nodes, edges, node types, edge types, respectively. Then,* $\alpha{:}\mathcal{V}{\rightarrow}\mathcal{A}$ *is a node type function that maps each node* $v{\in}\mathcal{V}$ *to one node type* $a{\in}\mathcal{A}$*, while* $\beta{:}\mathcal{E}{\rightarrow}\mathcal{R}$ *is an edge type function that maps each edge* $e{\in}\mathcal{E}$ *to an edge type* $r{\in}\mathcal{R}$*. An SKG is* heterogeneous *if it has multiple node and edge types such that* $|\mathcal{A}| + |\mathcal{R}| > 2$*.*

An SKG typically combines structural information (edges) with semantic details (nodes and their types). Its heterogeneous nature introduces complexity but also enables the modeling of diverse relationships and attributes among different entities. The nodes' attributes are the *metadata* – data providing information about other data – which are represented as key–value pairs describing properties such as the title, DOI, and description of the research outcomes.

Definition 2. *We denote the* ***publication*** *as* $\mathcal{V}_p$ *and the* ***dataset*** *nodes as* $\mathcal{V}_d$*, where* $\mathcal{V}_p \cap \mathcal{V}_d{=}\emptyset$*. The goal of* publication-dataset link prediction *task is to estimate the likelihood of an edge existing between two nodes* (u, v)*, where* $u \in \mathcal{V}_p$ *and* $v{\in}\mathcal{V}_d$*, such that* $\psi : (\mathcal{V}_p{\times}\mathcal{V}_d){\rightarrow}[0, 1]$*.*

In this work, we do not take into account the semantics associated with the edges. Although various semantic types exist to describe different kinds of relations between entities, we do not consider them in our analysis.

The publication-dataset link prediction task presents significant challenges due to the sparse and incomplete nature of SKGs, which often contain multiple connected components, duplicated nodes, missing links, and heterogeneous,

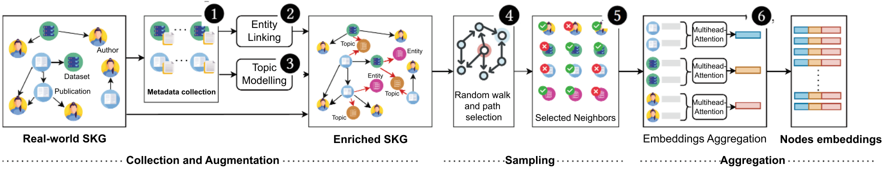


Fig. 1. The SAN Architecture comprising the collection and augmentation, neighbors sampling, and the aggregation phases.

incomplete metadata. Additionally, no existing methods are specifically designed to address link prediction in SKGs that exhibit these complexities. This challenge is further compunded by the absence of publicly available datasets tailored for this task. These factors underscore the problem's complexity and highlight the need for novel approaches to handle the structural intricacies and data gaps in SKGs.

We study the dataset link prediction task as a node-representation learning problem.

Definition 3. *Given an heterogeneous SKG $\mathcal{G}$, the objective of GRL is to learn a function $f\colon \mathcal{V} \rightarrow \mathbb{R}^d$ that maps each node $v{\in}\mathcal{V}$ to a d-dimensional embedding space, where $d{\ll}|\mathcal{V}|$. This embedding is designed such that for any two nodes $u \in \mathcal{V}_p$ and $v \in \mathcal{V}_d$, the dot-product of their corresponding vectors $\boldsymbol{u} \cdot \boldsymbol{v}$ approximates the scoring function ψ.*

GRL provides a way to map heterogeneous nodes into a shared embedding space, capturing their features and the graph's structure. This is crucial for the dataset link prediction task, where the relationship between query nodes (publications) and potential datasets must be effectively learned and represented in the embedding space.

4 Architecture

In this section, we present the SAN architecture, illustrated in Fig. 1. The pipeline consists of three phases: (1) collection and augmentation, (2) neighbor sampling, and (3) aggregation.

Collection and Augmentation. This phase is carried out offline to address the sparsity issue in SKGs. Given an SKG as input, SAN extracts the available textual metadata from the *publication* and *dataset* nodes (➊). Then, it performs entity linking (➋) [2] to identify and disambiguate entities within the textual content. Additionally, it applies topic modeling (➌) using BERTopic [14], which leverages transformer-based embeddings and c-TF-IDF to group similar documents based on shared topics.

The topics represent broader themes that act as central hubs within the SKG, connecting multiple disconnected components and improving the overall connectivity of the graph. On the other hand, entities are more specific and, when shared between nodes, suggest that those nodes are likely related by similar or closely related content. This distinction strengthens the graph's structure by using topics to boost general connectivity and entities to emphasize specific content overlaps between nodes. Entities and topics are subsequently integrated into the SKG as new nodes of type *entity* and *topic*. This phase results in an enhanced SKG. For each node, SAN extracts and concatenates two vectors: one based on graph topology using node2vec [15], and the other based on textual metadata using pre-trained models. In our case we employ the `all-MiniLM-L6-v2` sentence transformer[1] for longer textual metadata and `phrase-BERT` [33] for shorter text like topics. This phase aims to address the sparsity issue often affecting the SKGs, improving *robustness (RQ1)* by adding new nodes that enhance the graph's connectivity.

Sampling. Due to the small-world property of many graphs, multi-hop neighborhoods can include a substantial number of nodes. Therefore, given a target node v in the enriched SKG, this phase aims to explore the neighborhood of v and gather a representative set of neighbors of various types. A common neighbor sampling method, used in approaches such as GraphSAGE, is random sampling. However, this approach may produce an unrepresentative sample because SKGs are heterogeneous, with a highly imbalanced distribution of node types. In practice, the author nodes often lack disambiguation, leading to a large number of mostly isolated nodes in the SKG. In contrast, venue nodes are typically fewer and more interconnected.

Taking this issue into account, for a given target node v, we generate a set of random walks of length l originating from v (❹). We ignore edge directions in the SKG because for each edge we can assume there exists an edge in the opposite direction (e.g., for each `cites` edge we can assume a `cited by` exists in the opposite direction). Then, we select the k random walks with the highest similarity scores with the target node v, computed as follows:

$$\text{sim\_score}(v, \text{walk}_j) = \frac{1}{|\text{walk}_j|} \sum_{i \in \text{walk}_j} \frac{cos(\mathbf{v}_v, \mathbf{v}_i)}{d(v, i)}$$

where walk_j is one random walk, $m = |\text{walk}_j|$ is the number of publication and dataset nodes in the walk j, the node i is either a *publication* or a *dataset* ($i \in \{\mathcal{V}_\text{p} \cup \mathcal{V}_\text{d}\}$), $\mathbf{v}_v$ and $\mathbf{v}_i$ are the embeddings encoding the textual metadata of v and i respectively, $cos(\mathbf{v}_\text{v}, \mathbf{v}_i)$ is the cosine similarity, defined as $\frac{\mathbf{v}_\text{v} \cdot \mathbf{v}_i}{\|\mathbf{v}_\text{v}\| \|\mathbf{v}_i\|}$ and d is the distance – i.e., the number of nodes – between the target node v and the node i in the walk. In step (❺), SAN selects the top n publications and datasets with the highest cosine similarity to the target v from the k selected walks. For all other node types, it selects the n nodes closest to the target v within each type. In the sampling phase, the goal is to enhance *adaptability (RQ2)* by selecting a subset of representative neighbors instead of capturing the full neighborhood of

[1] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2.

a node. This strategy allows the model to adapt to dynamic graphs of varying sizes, without relying on global knowledge of the entire node set.

Aggregation. In the aggregation step (❻), SAN uses multihead attention [29] to combine neighbor vectors and derive the final embedding representation of the target node v. Multiple attention heads allow us to focus on different aspects of the input [29].

Formally, SAN combines three input matrices: Query ($\mathbf{Q}$), Key ($\mathbf{K}$), and Value ($\mathbf{V}$), of sizes (n_Q, d_{model}), (n_K, d_{model}), (n_V, d_{model}) where n_Q, n_K, n_V are the numbers of queries, keys, values embeddings respectively, and d_{model} their dimension. Queries, keys, and values are matrices constructed from node embeddings and used to compute attention scores. For each attention head $\vartheta_0, \ldots, \vartheta_\ell, \ldots, \vartheta_H$, these matrices are transformed as follows: $\mathbf{Q}_\ell{=}\mathbf{Q}\mathbf{W}_\ell^Q$, $\mathbf{K}_\ell{=}\mathbf{K}\mathbf{W}_\ell^K$, and $\mathbf{V}_\ell{=}\mathbf{V}\mathbf{W}_\ell^V$, where $\mathbf{W}_\ell^Q$, $\mathbf{W}_\ell^K$, and $\mathbf{W}_\ell^V$ are projection matrices of size (d_{model}, d_k) and $d_k{=}\frac{d_{model}}{H}$ is the dimensionality of the projection space for each head. For each head ϑ_ℓ, the attention scores are computed using the scaled dot-product attention mechanism:

$$\vartheta_\ell = \text{Attention}(\mathbf{Q}_\ell, \mathbf{K}_\ell, \mathbf{V}_\ell) = \text{softmax}\left(\frac{\mathbf{Q}_\ell \mathbf{K}_\ell^T}{\sqrt{d_k}}\right)\mathbf{V}_\ell.$$

The outputs from all heads are then concatenated and linearly projected to form the final output. The embeddings obtained are finally concatenated to obtain the final target node embedding representation.

We experimented with various embedding combination methods, including bi-LSTM, GRU, and mean pooling, both as replacements for multihead attention and for the final concatenation step. However, the combination of multihead attention and concatenation consistently outperformed these alternatives.

Multihead attention excels at capturing the intricate relationships between topological structures and textual metadata, allowing the model to understand how individual node features interact across different nodes and their descriptions. This capability is especially crucial in heterogeneous graphs, where different node and edge types may have varying levels of importance.

This phase enhances *versatility (RQ3)* by addressing the integration of multiple information sources, enabling the model to adjust to a range of real-world metadata scenarios.

Training. In this work, we used a mini-batch gradient descent training procedure. We optimized the model by minimizing the cross-entropy loss using negative sampling. Formally, the loss function is defined as:

$$\mathcal{L} = -\left(\sum_{(u,v)\in\mathcal{E}_{p,d}} \log\sigma\left(\mathbf{h}_u^\top \mathbf{h}_v\right) + \sum_{(u',v')\notin\mathcal{E}_{p,d}} \log\sigma\left(-\mathbf{h}_{u'}^\top \mathbf{h}_{v'}\right)\right),$$

where σ represents the sigmoid function, $\mathbf{h}_u$ and $\mathbf{h}_v$ are the embedding representations for publication u and dataset v, $\mathcal{E}_{\mathrm{p,d}}$ denotes the set of positive edges

Table 1. Nodes and edges statistics of the used datasets.

	Nodes		Edge			
MES	Publication (P)	2,1K	P–P	450	P–A	9,8K
	Dataset (D)	3K	P–D	2,5K	D–A	10,2K
	Author (A)	9,5K	D–D	1,2K		
PubMed	Publication (P)	42,6K	P–P	18,3K	D–K	17K
	Dataset (D)	33,8K	P–D	29,5K	P–V	42,6K
	Author (A)	334K	D–D	8K	P–O	87,5K
	Keyword (K)	10,8K	P–A	180K	D–O	515
	Venue (V)	6,7K	D–A	94,6K		
	Organization (O)	14,8K	P–K	6,6K		

between publications and datasets. For each positive edge $(u, v) \in \mathcal{E}_{p,d}$, we sample uniformly at random a negative pair $(u', v') \notin \mathcal{E}_{p,d}$, of publications and datasets not directly connected in the ground-truth.

5 Evaluation

We evaluate SAN across three increasingly complex settings. In the first setting, referred to as **transductive**, the entire set of publications and datasets is accessible during training, and the SKG remains unchanged. This standard setup has also been adopted in previous works [13,24,37]. In the second setting, **semi-inductive**, new publications are introduced at test time, while all datasets are available during training. The task involves predicting links between these unseen publications and the known datasets, and is relevant in order to connect to new publications datasets which can be may be thematically related. It is worth noting that this semi-inductive scenario is sometimes labeled as simply "inductive" in prior work [37]. In contrast, we define a **fully inductive** setting as one where both new publications and new datasets appear during testing, closely reflecting the dynamics of real-world SKGs. The goal here is to predict links between the new, unseen publications and datasets. Further, we test the models' robustness in the absence of textual metadata under the following different metadata conditions for each setting. The **ideal metadata setup** represents the ideal scenario where all datasets are fully described by complete metadata – i.e., all the datasets have a title and a description. The **real metadata setup** represents intermediate scenarios where only 75%, 50%, or 25% of the datasets have complete metadata, with the datasets containing metadata randomly selected. This setting reflects real-world conditions, as in real SKGs, datasets may or may not have comprehensive and descriptive metadata. The implementation of SAN and the related datasets are available on GitHub[2].

Datasets and Measures. We use two SKGs extracted from the OAG [23]: MES [19] and PubMed[3]. MES is a curated scientific graph interconnecting pub-

[2] https://github.com/HeterogeneousGL-SAN/SAN.

[3] Datasets are available at https://doi.org/10.6084/m9.figshare.26661382.v1.

lications, datasets, and authors; it represents a reliable snapshot of the scientific activity of the European Marine Science (MES) community of OpenAIRE. PubMed is a larger subgraph of the OAG also interconnecting a set of publications and datasets, authors, venues, organizations and keywords. Differently from MES, PubMed is not curated. The statistics of the resulting SKG are described in Table 1. We divided the $P-D$ (publication-dataset) edges into five sets: training, validation, and three test sets (transductive, semi-inductive, and fully inductive). The $P-D$ edges in the three test sets were randomly extracted from the original MES and PubMed datasets. Publications and datasets included in the semi-inductive and fully inductive test sets were removed from the original MES and PubMed datasets. From the filtered MES and PubMed datasets, we extracted a set of $P-D$ edges for validation and left the remaining part to train the model. In MES, we used $2K$ edges for training, 155 for validation and 155 for each of the three test sets; in PubMed we used $26.8K$ for training, $1.8K$ for validation, and $1.8K$ for each of the three test sets. For evaluation, we use the Area Under the ROC Curve (AUC) curve and F1-score. AUC and F1-score have been extensively used to evaluate link prediction performances of several graph learning approaches [13,37]. The AUC is a performance metric that measures a model's ability to distinguish between classes. The F1-score, on the other hand, is the harmonic mean of precision and recall. It provides a balanced measure that accounts for both false positives and false negatives, making it particularly useful in scenarios with class imbalance.

Parameters of SAN. We initialized publication and dataset nodes with embeddings from the `all-MiniLM-L6-v2` pretrained sentence transformer, chosen for its effectiveness on short text. For keywords, entities, and topics – typically short phrases of one to five words – we used `phrase-BERT` to generate their embeddings. Additionally, topological embeddings for all nodes were computed using `node2vec` [15]. Entity extraction was performed on publication and dataset metadata using DBpedia Spotlight [25], with a confidence threshold of 0.75. Topics were identified via BERTopic [14], using a minimum cluster size of 2 documents.

SAN was trained for 100 epochs with early stopping to reduce training time, using a learning rate of 10^{-5} and a mini-batch size of 1024. For each target node, we sampled 5 neighbors from publication or dataset nodes, 5 from keyword, topic, or entity types, and 5 from venue, author, and organization types. The multi-head attention mechanism used 8 heads and produced embeddings of dimension 128. Hyperparameters – including number of epochs, learning rate, batch size, attention heads, and neighbor sampling counts – were tuned via grid-search.

Baselines. We compared SAN with five inductive graph-based baselines. For each baseline, we report the best results obtained through grid search. The node embeddings were initialized by concatenating the textual embeddings using `all-MiniLM-L6-v2` with those from `node2Vec`. GraphSAGE [16] (SAGE for brevity) and GAT [30] were trained for 200 epochs with a learning rate of 10^{-5}

and early stopping. The output dimension was set to 128. Similarly, HAN [35] and HGT [17] were trained for 200 epochs with a learning rate of 10^{-4}, also utilizing early stopping, with an output dimension of 128. HetGNN [37] (HGNN for brevity) was trained for 100 epochs with early stopping, a learning rate of 10^{-5}, and 10 neighbors per node type, as specified in the original paper. To implement the baselines we used PyTorch Geometric [12].

Certain baselines were excluded from this study. GATNE [7] was not considered because it requires multiple relation types within the same set of nodes. MAGNN [13] was excluded since it was designed for the node classification task, and was not designed for the inductive settings. HiNormer [24] was also not included, as it has not been designed for link prediction tasks.

6 Results

In this section, we report the experimental results for dataset link prediction on MES and PubMed for all the tested settings.

Table 2 present the results for the link prediction task on the MES and PubMed datasets in the three settings: transductive, semi-inductive, and fully inductive. On the MES dataset, SAN consistently outperforms all baselines in nearly all scenarios both in terms of AUC and F1 score, except when only 25% of the datasets have no metadata where SAGE and GAT perform better than SAN in terms of F1 score. All models perform best in the transductive setting with full metadata availability, the default and least realistic scenario in the current academic landscape. Notably, in this scenario, the performance of all examined models, including SAN, tends to decrease slightly in semi- and fully inductive settings. This is reasonable as the prediction is made for unseen nodes. SAN and SAGE are the only methods whose AUC and F1 exceed the 0.9 in transductive, semi-inductive and inductive settings. The exception is HAN, which performs significantly worse in the semi-inductive and inductive settings. When evaluating performance in more realistic conditions where some datasets lack metadata, we observe that as the percentage of datasets with metadata decreases (from 75% to 25%), the performance of all the tested models decreases. This decline underscores the importance of text-based features for baselines. Furthermore, as the percentage of datasets with metadata decreases, the ability of the baselines to generalize to new, unseen data diminishes. SAN is the only method that consistently maintains an AUC above 0.9 and an F1 score of at least 0.8, even when the 50% of the dataset nodes lack textual metadata. This suggests that the SAN approach, combining different node embeddings while incorporating topology-based features, effectively mitigates the impact of missing text-based features, significantly affecting the performance of other baselines. Interestingly, SAGE and GAT, designed for homogeneous graphs, emerge as the most resilient baselines. In the MES dataset, reducing metadata coverage to 25% leads to a significant performance drop across all methods. For instance, SAN's AUC falls by 0.1 when metadata availability decreases from 50% to 25%. This is due to the limited number of datasets with metadata – only 750 – insufficient for accurate prediction in the MES SKG. Notably, SAN is robust when applied to real-world, sparse, large, and not curated SKGs like PubMed. As in the previous case, the performance of all tested methods declines in semi-inductive and inductive settings, along with the decreasing availability of textual metadata. This is expected as the setting becomes increasingly challenging. SAN demonstrates the greatest resilience to the absence of textual metadata and inductive settings, remaining the only method maintaining an AUC above 0.9 and an F1 score above 0.8 across all settings with more than 25% available metadata and 0.8 and 0.7, respectively, at 25%. In this regard, none of the other tested methods surpass an AUC of 0.871 (SAGE) or an F1 score of 0.818 (GAT). Furthermore, performance drops substantially as the number of datasets with available metadata decreases. In the most challenging scenario, with only 25% of datasets containing available metadata and within an inductive setup, SAN surpasses the best baseline (GAT) by 43% in AUC and 25% in F1. We see that HAN and HGNN are the models with the lowest performance on the PubMed dataset, having the AUC and F1

Table 2. AUC and F1 score over the MES (left) and PUBMED (right) datasets in transductive (Tran), semi-inductive (Semi), and inductive (Ind) settings. 100% indicates that all the datasets have textual metadata. $[75\%, 50\%, 25\%]$ indicates the percentage of datasets having textual metadata.

			MES						PUBMED					
			SAGE	GAT	HGT	HAN	HGNN	SAN	SAGE	GAT	HGT	HAN	HGNN	SAN
100%	Tran	AUC	0.980	0.933	0.921	0.838	0.912	**0.986**	0.871	0.825	0.839	0.578	0.752	**0.967**
		F1	0.958	0.927	0.860	0.825	0.871	**0.959**	0.798	0.818	0.766	0.532	0.700	**0.878**
	Semi	AUC	0.973	0.904	0.883	0.593	0.899	**0.983**	0.860	0.821	0.793	0.470	0.743	**0.948**
		F1	**0.956**	0.890	0.827	0.653	0.862	0.945	0.790	0.812	0.654	0.433	0.709	**0.860**
	Ind	AUC	0.945	0.896	0.772	0.494	0.872	**0.948**	0.854	0.818	0.710	0.453	0.722	**0.943**
		F1	0.920	0.880	0.746	0.326	0.821	**0.944**	0.790	0.809	0.523	0.422	0.701	**0.858**
75%	Tran	AUC	0.933	0.889	0.856	0.834	0.830	**0.971**	0.727	0.731	0.771	0.572	0.691	**0.942**
		F1	0.920	0.858	0.785	0.823	0.778	**0.926**	0.731	0.747	0.679	0.529	0.617	**0.859**
	Semi	AUC	0.920	0.883	0.845	0.592	0.821	**0.969**	0.707	0.719	0.732	0.469	0.632	**0.940**
		F1	0.885	0.860	0.759	0.652	0.783	**0.894**	0.714	0.735	0.665	0.411	0.609	**0.847**
	Ind	AUC	0.907	0.840	0.734	0.515	0.812	**0.957**	0.703	0.710	0.643	0.459	0.652	**0.938**
		F1	0.872	0.814	0.709	0.369	0.775	**0.876**	0.712	0.734	0.462	0.389	0.573	**0.846**
50%	Tran	AUC	0.901	0.867	0.801	0.815	0.806	**0.951**	0.612	0.644	0.703	0.568	0.567	**0.932**
		F1	0.808	0.817	0.702	0.800	0.754	**0.843**	0.652	0.663	0.668	0.489	0.502	**0.803**
	Semi	AUC	0.851	0.833	0.787	0.582	0.789	**0.944**	0.603	0.630	0.701	0.483	0.564	**0.922**
		F1	0.797	0.788	0.714	0.642	0.735	**0.840**	0.634	0.644	0.576	0.435	0.488	**0.809**
	Ind	AUC	0.753	0.802	0.691	0.505	0.776	**0.943**	0.601	0.627	0.584	0.478	0.499	**0.921**
		F1	0.746	0.769	0.684	0.326	0.710	**0.795**	0.632	0.640	0.532	0.412	0.464	**0.797**
25%	Tran	AUC	0.831	0.826	0.780	0.784	0.802	**0.840**	0.547	0.585	0.688	0.559	0.451	**0.900**
		F1	0.717	**0.745**	0.689	0.690	0.705	0.700	0.566	0.582	0.668	0.478	0.456	**0.716**
	Semi	AUC	0.826	0.771	0.725	0.529	0.764	**0.835**	0.541	0.565	0.649	0.473	0.433	**0.896**
		F1	0.726	**0.732**	0.664	0.629	0.700	0.688	0.564	0.553	0.451	0.421	0.438	**0.707**
	Ind	AUC	0.735	0.748	0.650	0.444	0.732	**0.790**	0.531	0.559	0.512	0.448	0.411	**0.800**
		F1	**0.646**	0.612	0.624	0.284	0.617	0.628	0.556	0.551	0.371	0.398	0.401	**0.693**

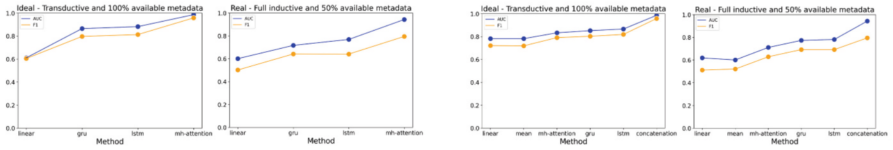
Table 3. AUC and F1 scores over the MES dataset in transductive and 100% metadata available setting. The column "no augmentation" refers to the models run excluding topics and entities; "augmentation" reports the results on the enriched graph.

Method	Augmentation		No Augmentation	
	AUC	F1	AUC	F1
GAT	0.872	0.831	0.933	0.927
SAGE	0.901	0.882	**0.980**	**0.958**
HGT	0.932	0.868	0.921	0.860
HAN	0.863	0.844	0.838	0.825
HGNN	0.933	0.884	0.912	0.871
SAN	**0.986**	**0.959**	0.912	0.876

always lower than 0.58 and 0.76, respectively, showing their low effectiveness on real-world SKGs.

7 Ablation Study

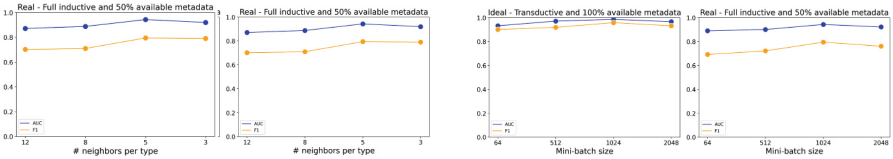
Components Analysis. We first assess how the augmentation phase – specifically, the integration of entity and topic nodes – affects performances. In particular, we investigate whether the enriched graph structure enables more meaningful neighborhood interactions and improves the model's ability to capture relevant relationships. In Table 3 presents the performance of five systems with ("augmentation" column) and without ("no augmentation" column) this phase. For GAT and GraphSAGE, the inclusion of heterogeneous node types reduces model effectiveness. In SKGs, embeddings are generated differently depending on node type, and aggregating such diverse representations can hinder overall performance. In contrast, SAN and HGNN gain from the introduction of additional nodes, showing improved results on the enriched graph. This highlights that augmentation is especially advantageous for models designed to leverage structural properties and heterogeneity. For HAN and HGT, omitting the augmentation phase results in a slight performance decrease. In conclusion, augmentation benefits models like SAN and HGNN that leverage heterogeneity and graph structure, while it may have limited or negative effects on models not designed for diverse node types. To assess the effectiveness of different aggregation strategies within SAN, we compared multihead attention, bi-LSTM, GRU, and linear projection. The evaluation was based on AUC and F1 scores on the MES dataset across two settings: an ideal transductive scenario with 100% metadata availability, and a more realistic fully-inductive scenario with only 50% metadata available. Figure 2a shows that multihead attention demonstrates the highest robustness, with bi-LSTM and GRU following closely behind. In contrast, linear projection results in the lowest performance in the dataset link prediction task. These findings suggest that multihead attention is particularly well-suited, as it allows the model to focus on different parts of the embedding space, capturing diverse relational patterns and improving overall predictive performance. Moreover, in scenarios where textual metadata is limited, multihead attention can better leverage topological information, compensating for the lack of text-based features. Finally, we



(a) Component-based analysis of the aggregation phase.

(b) Component-based analysis of the combination phase.

Fig. 2. Comparison of aggregation (on the left) and combination phases in the pipeline (on the right).



(a) Number of neighbors of a target node sampled for each node type.

(b) Mini-batch size analysis.

Fig. 3. Hyperparameter analysis: number of neighbors (on the left), and mini-batch size (on the right).

assessed the performance of SAN by experimenting with alternative embedding combination strategies in place of simple concatenation. While SAN originally relies on concatenation to merge embeddings, we evaluated several other methods, including bi-LSTM, GRU, multihead attention, linear projection, and mean pooling. Figure 2b presents the results on the MES dataset under both ideal and real-world conditions. Among the tested approaches, concatenation remains the most effective, whereas GRU and linear projection yield the weakest performance.

Hyperparameter Analysis. One key factor that affects efficiency, is the number of neighbors sampled for each node type when targeting a node v_t. SKGs tend to have an imbalanced distribution of node types, so selecting the most representative nodes is crucial to ensure a well-balanced set. Sampling too many nodes may introduce noise by including irrelevant or redundant information, while sampling too few could lead to an incomplete and unrepresentative set of neighbors. Another critical aspect influencing the overall efficiency of the model is the mini-batch size. Smaller batches enable faster training but increase the likelihood of overfitting. On the other hand, larger batches may reduce the risk of overfitting, yet they can lead to slower convergence and may require more computational resources. Figures 3a and 3b display our analysis on the MES dataset. We compared two scenarios: the ideal and the real-world setting. In both cases, performance shows slight variations with the number of neighbors considered, reaching a peak when selecting five neighbors per type. These small fluctuations suggest that multihead attention effectively mitigates the effects of additional or missing nodes, highlighting its adaptability and robustness in real-world SKGs. Regarding mini-batch size, we observe that batches smaller than 1024 lead to a higher risk of overfitting, while larger batches tend to result in slower convergence and increased computational requirements.

8 Conclusions

Scientific data is crucial for research progress, yet datasets are often poorly described and hard to find. While curated SKGs focus mostly on publications, those that include datasets tend to be large but uncurated, with incomplete metadata and weak interlinking. This results in SKGs that are noisy, sparse, and heterogeneous, making data discovery and citation difficult. In this context, link prediction becomes an essential task.

We propose SAN, a method for enriching SKGs and performing heterogeneous GRL. SAN adopts a three-phase approach: it enhances graph connectivity using text-derived nodes, selects relevant neighbors via random walks, and combines their representations with multihead attention. This architecture seamlessly integrates textual information with graph topology while not relying on complete metadata coverage. Additionally, we conduct extensive experiments across three settings (transductive, semi-inductive, fully inductive) and under two metadata conditions (ideal and real) using two benchmark datasets. Our evaluation emphasizes the often overlooked importance of generating predictions for newly added items without requiring model retraining. The results demonstrate SAN's effectiveness compared to five inductive baselines. We explored three research questions focused on **robustness**, **adaptability**, and **versatility**. Our findings show that SAN delivers strong publication-dataset link prediction performances even in inductive scenarios with little textual metadata available, thanks to its ability to extract and utilize topology-based features. Current methods struggle to effectively use high-quality textual metadata. As a consequence, future research should explore more advanced architectures to address this gap. Further experiments should use SAN for other link prediction tasks, like author-venue and publication-venue predictions, to encourage collaborations and suggest relevant venues.

Acknowledgments. This work is supported by the HEREDITARY Project, as part of the European Union's Horizon Europe research and innovation programme under Grant Agreement No GA 101137074.

References

1. Akujuobi, U., Zhang, X.: Delve: A dataset-driven scholarly search and analysis system. SIGKDD Explor. **19**(2), 36–46 (2017). https://doi.org/10.1145/3166054.3166059
2. Balog, K.: Entity-oriented search, The Information Retrieval Series, vol. 39. Springer (2018). https://doi.org/10.1007/978-3-319-93935-3
3. Bordes, A., Usunier, N., García-Durán, A., Weston, J., Yakhnenko, O.: Translating embeddings for modeling multi-relational data. In: Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States, pp. 2787–2795 (2013), https://proceedings.neurips.cc/paper/2013/hash/1cecc7a77928ca8133fa24680a88d2f9-Abstract.html
4. Brickley, D., Burgess, M., Noy, N.: Google dataset search: building a search engine for datasets in an open web ecosystem. In: Proceedings of the ACM Web Conference 2019, WWW, pp. 1365–1375. ACM (2019). https://doi.org/10.1145/3308558.3313685, https://doi.org/10.1145/3308558.3313685
5. Buneman, P., et al.: Why data citation isn't working, and what to do about it. Database **2020** (2020)
6. Buneman, P., Dosso, D., Lissandrini, M., Silvello, G.: Data citation and the citation graph. Quant. Sci. Stud. **2**(4), 1399–1422 (2021). https://doi.org/10.1162/qss_a_00166
7. Cen, Y., Zou, X., Zhang, J., Yang, H., Zhou, J., Tang, J.: Representation learning for attributed multiplex heterogeneous network. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD, pp. 1358–1368. ACM (2019). https://doi.org/10.1145/3292500.3330964
8. Chen, J., Hou, H., Gao, J., Ji, Y., Bai, T.: Rgcn: recurrent graph convolutional networks for target-dependent sentiment analysis. In: Knowledge Science, Engineering and Management - 12th International Conference, KSEM 2019, Proceedings, Part I. Lecture Notes in Computer Science, vol. 11775, pp. 667–675. Springer (2019). https://doi.org/10.1007/978-3-030-29551-6_59
9. Costello, M.J., Michener, W.K., Gahegan, M., Zhang, Z.Q., Bourne, P.E.: Biodiversity data should be published, cited, and peer reviewed. Trends Ecol. Evol. **28**(8), 454–461 (2013)
10. Dong, Y., Chawla, N.V., Swami, A.: metapath2vec: scalable representation learning for heterogeneous networks. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 135–144. ACM, New York, NY, USA (2017). https://doi.org/10.1145/3097983.3098036
11. Färber, M., Lamprecht, D.: The data set knowledge graph: creating a linked open data source for data sets. Quant. Sci. Stud. **2**(4), 1324–1355 (2021)
12. Fey, M., Lenssen, J.E.: Fast graph representation learning with pytorch geometric. arXiv preprint arXiv:1903.02428 (2019)
13. Fu, X., Zhang, J., Meng, Z., King, I.: Magnn: Metapath aggregated graph neural network for heterogeneous graph embedding. In: Proceedings of the ACM Web Conference 2020, WWW, pp. 2331–2341. ACM, New York, NY, USA (2020). https://doi.org/10.1145/3366423.3380297
14. Grootendorst, M.: Bertopic: neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794 (2022)

15. Grover, A., Leskovec, J.: node2vec: scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 855–864. ACM (2016). https://doi.org/10.1145/2939672.2939754
16. Hamilton, W., Ying, Z., Leskovec, J.: Inductive representation learning on large graphs. In: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, pp. 1024–1034. Curran Associates, Inc. (2017)
17. Hu, Z., Dong, Y., Wang, K., Sun, Y.: Heterogeneous graph transformer. In: Proceedings of the ACM Web Conference 2020, WWW, pp. 2704–2710. ACM (2020). https://doi.org/10.1145/3366423.3380027
18. Irrera, O., Lissandrini, M., Dell'Aglio, D., Silvello, G.: Reproducibility and analysis of scientific dataset recommendation methods. In: Proceedings of the 18th ACM Conference on Recommender Systems, RecSys 2024, Bari, Italy, October 14-18, 2024, pp. 570–579. ACM (2024). https://doi.org/10.1145/3640457.3688071
19. Irrera, O., Mannocci, A., Manghi, P., Silvello, G.: A novel curated scholarly graph connecting textual and data publications. J. Data Inform. Qual. **15**(3) (2023). https://doi.org/10.1145/3597310
20. Jaradeh, M.Y. et al.: Open research knowledge graph: Next generation infrastructure for semantic scholarly knowledge. In: Proceedings of the 10th International Conference on Knowledge Capture, K-CAP 2019, Marina Del Rey, CA, USA, November 19-21, 2019 pp. 243–246. ACM (2019). https://doi.org/10.1145/3360901.3364435
21. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
22. Lv, Q., et al.: Are we really making much progress? revisiting, benchmarking and refining heterogeneous graph neural networks. In: KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 1150–1160. ACM (2021). https://doi.org/10.1145/3447548.3467350
23. Manghi, P., et al.: Openaire research graph dump (Dec 2022). https://doi.org/10.5281/zenodo.7488618. A new version of this dataset is published every 6 months. The content available on the OpenAIRE EXPLORE and CONNECT portals might be more up-to- date with respect to the data you find here
24. Mao, Q., Liu, Z., Liu, C., Sun, J.: Hinormer: Representation learning on heterogeneous information networks with graph transformer. In: Proceedings of the ACM Web Conference 2023, WWW, pp. 599–610. ACM (2023). https://doi.org/10.1145/3543507.3583493
25. Mendes, P.N., Jakob, M., García-Silva, A., Bizer, C.: Dbpedia spotlight: shedding light on the web of documents. In: Proceedings the 7th International Conference on Semantic Systems, I-SEMANTICS, pp. 1–8 (2011)
26. Silvello, G.: Theory and practice of data citation. J. Am. Soc. Inf. Sci. **69**(1), 6–20 (2018)
27. Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., Su, Z.: Arnetminer: extraction and mining of academic social networks. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 990–998. ACM, New York, NY, USA (2008). https://doi.org/10.1145/1401890.1402008
28. Trouillon, T., Welbl, J., Riedel, S., Gaussier, E., Bouchard, G.: Complex embeddings for simple link prediction. In: Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016. JMLR Workshop and Conference Proceedings, vol. 48, pp. 2071–2080. JMLR.org (2016). http://proceedings.mlr.press/v48/trouillon16.html

29. Vaswani, A.: Attention is all you need. arXiv preprint arXiv:1706.03762 (2017)
30. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y.: Graph attention networks. arXiv preprint arXiv:1710.10903 (2017)
31. Viswanathan, V., Gao, L., Wu, T., Liu, P., Neubig, G.: DataFinder: Scientific dataset recommendation from natural language descriptions. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers, pp. 10288–10303. ACL (2023). https://doi.org/10.18653/v1/2023.acl-long.573, https://aclanthology.org/2023.acl-long.573
32. Wang, K., Shen, Z., Huang, C., Wu, C., Dong, Y., Kanakia, A.: Microsoft academic graph: when experts are not enough. Quant. Sci. Stud. **1**(1), 396–413 (2020). https://doi.org/10.1162/qss_a_00021
33. Wang, S., Thompson, L., Iyyer, M.: Phrase-BERT: improved phrase embeddings from BERT with an application to corpus exploration. In: Proc. of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 10837–10851. ACL (2021)
34. Wang, X., Bo, D., Shi, C., Fan, S., Ye, Y., Philip, S.Y.: A survey on heterogeneous graph embedding: methods, techniques, applications and sources. IEEE Trans. Big Data **9**(2), 415–436 (2023). https://doi.org/10.1109/TBDATA.2022.3177455
35. Wang, X., et al.: Heterogeneous graph attention network. In: Proceedings of the ACM Web Conference 2019, WWW, pp. 2022–2032. ACM, New York, NY, USA (2019). https://doi.org/10.1145/3308558.3313562
36. Wilkinson, M.D., et al.: The fair guiding principles for scientific data management and stewardship. Sci. Data **3**(1), 160018 (Mar 2016). https://doi.org/10.1038/sdata.2016.18
37. Zhang, C., Song, D., Huang, C., Swami, A., Chawla, N.V.: Heterogeneous graph neural network. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 793–803. ACM (2019). https://doi.org/10.1145/3292500.3330961