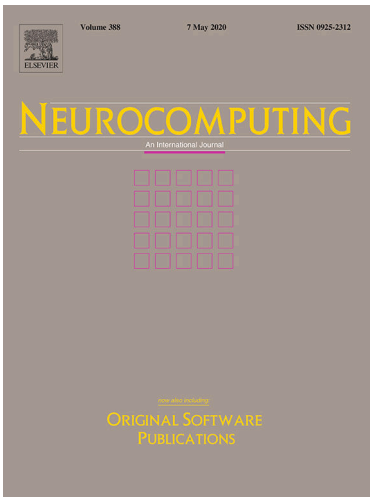
Journal Pre-proofs



Trans4E: Link Prediction on Scholarly Knowledge Graphs

Mojtaba Nayyeri, Gokce Muge Cil, Sahar Vahdati, Francesco Osborne, Mahfuzur Rahman, Simone Angioni, Angelo Salatino, Diego Reforgiato Recupero, Nadezhda Vassilyeva, Enrico Motta, Jens Lehmann

PII: S0925-2312(21)00960-7
DOI: https://doi.org/10.1016/j.neucom.2021.02.100
Reference: NEUCOM 23989

To appear in: *Neurocomputing*

Accepted Date: 12 February 2021

Please cite this article as: M. Nayyeri, G.M. Cil, S. Vahdati, F. Osborne, M. Rahman, S. Angioni, A. Salatino, D.R. Recupero, N. Vassilyeva, E. Motta, J. Lehmann, Trans4E: Link Prediction on Scholarly Knowledge Graphs, *Neurocomputing* (2021), doi: https://doi.org/10.1016/j.neucom.2021.02.100

Trans4E: Link Prediction on Scholarly Knowledge Graphs

Mojtaba Nayyeri[a], Gokce Muge Cil[a], Sahar Vahdati[b], Francesco Osborne[d], Mahfuzur Rahman[a], Simone Angioni[e], Angelo Salatino[d,*], Diego Reforgiato Recupero[e], Nadezhda Vassilyeva[a], Enrico Motta[d] and Jens Lehmann[a,c]

[a]SDA Research Group, University of Bonn (Germany)
[b]Institute for Applied Informatics (InfAI)
[c]Fraunhofer IAIS, Dresden (Germany)
[d]Knowledge Media Institute, The Open University, Milton Keynes (UK)
[e]Department of Mathematics and Computer Science, University of Cagliari (Italy)

ARTICLE INFO

Keywords:
Scholarly Knowledge Graph
Knowledge Graph Embedding
Scholarly Communication
Science Graph
Metaresearch Queries
Link Prediction
Research of Research

ABSTRACT

The incompleteness of Knowledge Graphs (KGs) is a crucial issue affecting the quality of AI-based services. In the scholarly domain, KGs describing research publications typically lack important information, hindering our ability to analyse and predict research dynamics. In recent years, link prediction approaches based on Knowledge Graph Embedding models became the first aid for this issue. In this work, we present Trans4E, a novel embedding model that is particularly fit for KGs which include N to M relations with N≫M. This is typical for KGs that categorize a large number of entities (e.g., research articles, patents, persons) according to a relatively small set of categories. Trans4E was applied on two large-scale knowledge graphs, the Academia/Industry DynAmics (AIDA) and Microsoft Academic Graph (MAG), for completing the information about Fields of Study (e.g., 'neural networks', 'machine learning', 'artificial intelligence'), and affiliation types (e.g., 'education', 'company', 'government'), improving the scope and accuracy of the resulting data. We evaluated our approach against alternative solutions on AIDA, MAG, and four other benchmarks (FB15k, FB15k-237, WN18, and WN18RR). Trans4E outperforms the other models when using low embedding dimensions and obtains competitive results in high dimensions.

1. Introduction

The technology of Knowledge Graphs (KGs) empowered by graph-based knowledge representation brought an evolutionary change in a range of AI tasks. As a consequence, many application domains in science, industry, and different enterprises use KGs for data management. However, a challenge with KGs is that, despite the presence of millions of triples, capturing complete knowledge from the real world is almost impossible, even for specific application domains. Therefore, KGs usually remain incomplete.

Scientific research is one of the major domains for the application of KGs. In the last years, we saw the emergence of several KGs describing research outputs, such as Microsoft Academic Graph[1] [51], Scopus[2], Semantic Scholar[3], Aminer [62], Core [21], OpenCitations [37], Dimensions[4], Open Research Knowledge Graph[5] [18], and others. These solutions are crucial for performing large-scale bibliometric studies, informing funding agencies and research policymakers, supporting a variety of intelligent systems for querying the scientific literature, identifying research topics, suggesting relevant articles and experts, detecting research trends, and so on. Their usefulness and, consequently, our ability to assess research dynamics, are however crucially limited by their incompleteness. Even basic metadata such as affiliations, organization types, references, research topics, and conferences are often missing, noisy, or not properly disambiguated. Therefore, apparently simple tasks such as identifying the affiliation and the country of origin of a publication still require a large amount of manual data cleaning [27].

Traditionally, data integration methods have been applied to solve data incompleteness in the context of databases and repositories. However, when completing and refining large KGs, it is crucial to adopt scalable and automatic approaches. Among the many possible graph completion methods, Knowledge Graph Embedding (KGE) models have recently gained a lot of attention. KGEs learn representations of graph nodes and edges with the goal of predicting links between existing entities. Embedding models have been in practical use for various types of KGs in different domains, including digital libraries [58], biomedical [25], and social media [43].

However, the specific characteristics of scholarly KGs poses important challenges for link prediction methods

*Corresponding author
nayyeri@cs.uni-bonn.de (M. Nayyeri); s6gocill@uni-bonn.de (G.M. Cil); vahdati@infai.org (S. Vahdati); francesco.osborne@open.ac.uk (F. Osborne); s6mrrahm@uni-bonn.de (M. Rahman); simone.angioni@unica.it (S. Angioni); angelo.salatino@open.ac.uk (A. Salatino); diego.reforgiato@unica.it (D.R. Recupero); vassilyeva@cs.uni-bonn.de (N. Vassilyeva); enrico.motta@open.ac.uk (E. Motta); jens.lehmann@cs.uni-bonn.de (J. Lehmann)
ORCID(s): 0000-0000-0000-0000 (M. Nayyeri); 0000-0002-4645-0088 (G.M. Cil); 0000-0002-7171-169X (S. Vahdati); 0000-0001-6557-3131 (F. Osborne); 0000-0003-0273-3112 (M. Rahman); 0000-0002-6682-3419 (S. Angioni); 0000-0002-4763-3943 (A. Salatino); 0000-0001-8646-6183 (D.R. Recupero); 0000-0000-0000-0000 (N. Vassilyeva); 0000-0003-0015-1952 (E. Motta); 0000-0000-0000-0000 (J. Lehmann)

[1]Microsoft Academic Graph - http://aka.ms/microsoft-academic
[2]Scopus - https://www.scopus.com/
[3]Semantic Scholar - https://www.semanticscholar.org/
[4]Dimensions - https://www.dimensions.ai/
[5]ORKG - https://www.orkg.org/orkg/

based on KGE models [7, 45, 48, 61, 53, 47, 29]. One crucial aspect is the presence of several N to M relations with N≫M. Given a triple (h, r, t), this situation arises when the cardinality of the entities in the head position (h) for a certain relation (r) is much higher than the one of the entities in the tail position (t). This is the case for most scholarly knowledge graphs [37, 51, 1, 62, 20] that usually categorize millions of documents (e.g., papers, patents) according to a relatively small set of categories (e.g., topics, affiliation kinds, countries, chemical compounds). Current KGE models lack the ability to handle effectively these kinds of relations since they are unable to assign to each entity a well distinct embedding vector in a low dimensional space. As a result, link prediction and node classification techniques that exploit these embeddings tend to perform poorly.

To address this problem, we propose Trans4E, a new embedding model specifically designed to support link prediction for KGs which present N to M relations with N≫M. Specifically, Trans4E tackles the issue by providing a larger number of possible vectors ($8^d - 1$, where d is the embedding dimension) to be assigned to entities involved in N to M relations. Trans4E enables the generation of a well distinct vector for each entity even when using small embedding dimensions.

The motivating scenario for this work was supplied by the Academia/Industry DynAmics (AIDA)[6] Knowledge Graph [2], a resource that was designed for studying the relationship between academia and industry and for supporting systems for predicting research dynamics. The current version of AIDA integrates the metadata about 21M publications from Microsoft Academic Graph (MAG) and 8M patents from Dimensions in the field of Computer Science. In this resource, documents are categorized according to their research topics drawn from the Computer Science Ontology (CSO)[7] [39] and classified with their authors' affiliation types on the Global Research Identifier Database (GRID)[8] (e.g., 'Education', 'Company', 'Government', 'Nonprofit'). This solution enables analysing the evolution of research topics across academia, industry, government institutions, and other organizations. For instance, it allows us to detect that a specific topic, originally introduced by academia, has been recently adopted by industry. It can also support systems for predicting the impact of specific research efforts on the industrial sector [38] and the evolution of technologies [34]. Nevertheless, only 5.1M out of the 21M articles could be mapped to a GRID and characterized according to their affiliation type. Therefore, more than 75% of the publications are missing this critical information, significantly reducing the scope and accuracy of the resulting analytics. In order to show that our approach can be applied to fields with very different characteristics, we also use it to complete the Fields of Study, which is a collection of terms from multiple disciplines utilized to index the articles in MAG. Indeed, the completeness of the set of terms associated with a paper varies a lot and depends on the quality and style of the abstract, which in turn is often parsed from online PDFs, leading to mistakes and missing content. This in turn hinders our ability to understand the research concepts associated with the paper and to obtain comprehensive analytics. Completing the affiliation types and the Fields of Study is crucial for improving the overall quality of these knowledge graphs and a very good practical use case for link prediction.

We evaluated Trans4E against several alternative models (TransE, RotatE, QuatE, ComplEx) on the task of link prediction on AIDA, MAG, and four other well-known benchmarks (FB15K, FB15k-237, WN18, and WN18RR).

The experiments showed that Trans4E outperforms the other approaches in the case of N to M relations with N ≫ M and yields very competitive results in all the other cases, in particular when using low embedding dimensions. The ability to solve the N ≫ M issue and to perform well even when adopting small embedding dimensions makes Trans4E particularly apt for handling large scale knowledge graphs that describe millions of entities of the same type (e.g., documents, persons).

In summary, the contributions of our work are the following:

- We propose Trans4E, a new embedding model specifically designed to provide link prediction for large-scale KGs presenting N to M relations with N ≫ M.
- We apply Trans4E on a real word scenario that involves completing affiliation types and Fields of Study (N ≫ M relations) in AIDA and MAG.
- We further evaluate our approach on four well-known benchmarks (FB15k, FB15k-237, WN18, and WN18RR), showing that Trans4E yields competitive performances in several configurations.

The rest of the paper is organised as follows. In Section 2, we review the literature on current embedding models for data completion and scholarly knowledge graphs. In Section 3, we present a motivating scenario involving the completion of the AIDA knowledge graph. In Section 4, we describe Trans4E. Section 5 reports the evaluation of the model versus alternative solutions. Finally, in Section 6 we summarise the main conclusions and outline future directions of research.

2. Background and Related Work

In this section, we will first review the graph embedding models and their application to link prediction. Then we will discuss the current generation of scholarly knowledge graphs that can benefit from these solutions.

2.1. Knowledge Graph Embedding Models

In this section, we introduce the definitions required to understand our approach.

[6] AIDA - http://aida.kmi.open.ac.uk
[7] CSO - https://cso.kmi.open.ac.uk/
[8] GRID - https://www.grid.ac/

Embedding Vectors. Let the knowledge graph be $KG = (\mathcal{E}, \mathcal{R}, \mathcal{T})$, where $\mathcal{E}$ is the set of entities (nodes) in the graph, $\mathcal{R}$ is the set of all relations (edges), and $\mathcal{T}$ is the set of all triples in the graph in the form of (h, r, t), e.g., *(Berlin, CapitalOf, Germany)*. KGE models are applied to KGs for link prediction by measuring the degree of correctness of a triple. To do so, a KGE model aims at mapping each entity and relation of the graph into a vector space (shown as $(\mathbf{h}, \mathbf{r}, \mathbf{t})$, $\mathbf{h}, \mathbf{r}, \mathbf{t} \in \mathbb{R}^d$), where d is the embedding dimension of each vector. By h_i, we refer to the i-th element of the vector $\mathbf{h}$ where i ranges in $\{1, \ldots, d\}$. The vector representation of the entities and the relations in a KG are the actual embeddings.

Score Function. Using this representation, the plausibility of the triples is then assessed by the scoring function $f(\mathbf{h}, \mathbf{r}, \mathbf{t})$ of the applied KGE model. If a triple is more plausible, its score should be higher. For example, f(*Berlin, CapitalOf, Germany*) should be higher than f(*Berlin, CapitalOf, France*).

Negative Sampling. Typical machine learning approaches are trained on both positive and negative samples. However, all the triples present in KGs are considered true, and this necessitates the injection of negative samples in the training of the KGEs. In this work we use Adversarial negative sampling (*adv*) for this purpose. This technique generates a set of negative samples from a triple (h, r, t) by using a probabilistic algorithm to replace h or t with a random entity (h' or t') existing in $\mathcal{E}$.

Loss Function. Since at the beginning of the learning process, the embedding vectors are initialized with random values, the scores of the triples for positive and negative samples are also random. Optimization of a loss function $\mathcal{L}$ is utilized to adjust the embeddings in such a way that positive samples get higher scores than the negatives ones. Stochastic Gradient Descent (SGD) method is commonly used for optimising the loss function.

N to M Relations. As mentioned above, given a relation r, the representation of facts in triple form is (h, r, t). Depending on the type of a relation and its meaning, for a fixed head (say h_1), there are at most M possible tails connected to the head, i.e. $\{(h_1, r, t_1), (h_1, r, t_2)\}, \ldots, (h_1, r, t_M)$. Similarly, for a fixed tail, (say t_1), there are at most N possible head entity, i.e. $\{(h_1, r, t_1), (h_2, r, t_1), \ldots, (h_N, r, t_1)\}$. There are four cases that may arise for a relations which connects a different number of heads and tails: a) both N and M are small, b) both M and N are large, c) N is small and M is large, and d) N is large and M is small. The latter is the focus of this paper. For example, in the AIDA knowledge graph the "hasType" relations connects a very large number of head entities (5.1M articles) to only 8 tail entities (the GRID types).

2.2. Review of State-of-the-art KGEs

Here we summarize some of the most used existing models focusing in particular on their scoring function.

TransE [7] is one of the early embedding models and is well known for its outstanding performance and simplicity. It is a solid baseline that can still outperform many of the most recent and complex KGEs [17]. The idea of the TransE model is to enforce embedding of entities and relations in a positive triple (h, r, t) to satisfy the following equality:

$$\mathbf{h} + \mathbf{r} \approx \mathbf{t} \tag{1}$$

where $\mathbf{h}$, $\mathbf{r}$ and $\mathbf{t}$ are the embedding vectors of head, relation, and tail, respectively. TransE model defines the following scoring function:

$$f_r(h, t) = -\|\mathbf{h} + \mathbf{r} - \mathbf{t}\| \tag{2}$$

RotatE [45] is a model designed to transform the head entity to the tail entity by using the relation rotation. This model embeds entities and relations in complex space. If we constrain the norm of entity vectors, this model would be reduced to TransE. The scoring function of RotatE is

$$f_r(h, t) = -\|\mathbf{h} \circ \mathbf{r} - \mathbf{t}\| \tag{3}$$

in which $\circ$ is the element-wise product. Rotate is one of the recent state-of-the-art models which is leading the accuracy competition among KGEs [45].

ComplEx [48] is a semantic matching model, which assesses the plausibility of facts by considering the similarity of their latent representations. In other words, it is assumed that similar entities have common characteristics, i.e. are connected through similar relationships [32, 52]. In ComplEx the entities are embedded in the complex space. The score function of ComplEx is given as follows:

$$f(h, t) = \Re(\mathbf{h}^T \operatorname{diag}(\mathbf{r})\, \bar{\mathbf{t}})$$

in which $\bar{\mathbf{t}}$ is the conjugate of the vector $\mathbf{t}$ and $\Re$ returns the real part of the complex number.

QuatE [61] models relations in the quaternion space. Similarly to RotatE, QuatE represents a relation as a rotation. However, a rotation in quaternion space is more expressive than a rotation in complex space. A product of two quaternions $Q_1 \otimes Q_2$ is equivalent to first scaling Q_1 by magnitude $|Q_2|$ and then rotating it in four dimensions. QuatE finds a mapping $\mathcal{E} \rightarrow \mathbb{H}^d$, where an entity h is represented by a quaternion vector $\mathbf{h} = a_h + b_h\mathbf{i} + c_h\mathbf{j} + d_h\mathbf{k}$, with $a_h, b_h, c_h, d_h \in \mathbb{R}^d$.

The scoring function is computed as follows:

$$\phi(h, r, t) = \mathbf{h}' \cdot \mathbf{t} = \langle a'_h, a_t \rangle + \langle b'_h, b_t \rangle + \langle c'_h, c_t \rangle + \langle d'_h, d_t \rangle \tag{4}$$

where $\langle \cdot, \cdot \rangle$ is the inner product. $\mathbf{h}'$ is computed by first normalizing the relation embedding $\mathbf{r} = p_r + q_r\mathbf{i} + u_r\mathbf{j} + v_r\mathbf{k}$ to a unit quaternion:

$$\mathbf{r}^{(n)} = \frac{r}{|r|} = \frac{p_r + q_r\mathbf{i} + u_r\mathbf{j} + v_r\mathbf{k}}{\sqrt{p_r^2 + q_r^2 + u_r^2 + v_r^2}} \tag{5}$$

and then computing the Hamiltonian product between $\mathbf{r}^{(n)}$ and $\mathbf{h} = a_h + b_h\mathbf{i} + c_h\mathbf{j} + d_h\mathbf{k}$:

$$\begin{aligned}\mathbf{h}' = \mathbf{h} \otimes \mathbf{r}^{(n)} := &\left(a_h \circ p - b_h \circ q - c_h \circ u - d_h \circ v\right) \\ &+ \left(a_h \circ q + b_h \circ p + c_h \circ v - d_h \circ u\right)\mathbf{i} \\ &+ \left(a_h \circ u - b_h \circ v + c_h \circ p + d_h \circ q\right)\mathbf{j} \\ &+ \left(a_h \circ v + b_h \circ u - c_h \circ q + d_h \circ p\right)\mathbf{k}\end{aligned} \tag{6}$$

2.3. Further Related Work

Beside KGE models such as QuatE, ComplEx, TransE, and RotatE, there are several related approaches based on neural networks for KG completion. Here we cover the most relevant ones with specific focus on link prediction.

Few-Shot Learning (FSL) takes advantage of prior knowledge for efficiently learning from a limited number of examples [54, 59]. Some FSL methods focus on graph meta-learning [5, 60, 8], which provide fast adaption to the newly imported data. Such techniques are not reported to be suitable for large scale KGs and are mainly used on image datasets. Some other FSL methods focus on link prediction on graphs. For instance, authors in [10] use meta information for learning the most important and relevant knowledge with high retrieval speed. This model is similar to TransE and therefore suffers from the same limitation regarding N to M relation types with N $\gg$ M. In addition, FSL is most useful in scenarios with few examples. This is not the case of large-scale scholarly knowledge graph that usually can produce a large number of examples.

Transfer Learning (TL) aims at reusing the knowledge gained while solving a specific problem for solving a different but related one [64]. This strategy enables to address new learning tasks without extensive re-training [14, 35]. A family of methods labelled Graph Transfer Learning (GTLs) are specifically designed to work on graph-structured data [24]. However, they are not directly applicable to the task discussed in this paper since they mostly focus on similarity between entities rather than link prediction.

Graph Neural Networks (GNNs) are often used for link prediction [3, 57]. These methods compute the state of embeddings for a node according to the local neighborhood [63, 55, 50]. The embedding of a node is then created as a d-dimension vector and produces output information such as the node label. However, the high computation costs of GNNs make them unsuitable for large-scale knowledge graphs.

Several approaches for link prediction based on Deep Neural Networks, such as KBAT [28] and CapsE [49], reported very high performance, but were not consistent across different benchmarks. An analysis by Su at al. [46] showed that this behaviour was due to an inappropriate evaluation protocol, and the performance of these models dropped after fixing the relevant biases. According to them, instead, shallow KGE models (e.g., TransE, RotatE, ComplEx, QuatE) are able to perform consistently across several evaluation protocols [46].

Another interesting family of approaches regards community embeddings, which can optimize node embedding by using a community-aware high-order proximity [9]. For instance, vGraph [44] is a probabilistic generative model that learns community membership and node representation collaboratively. ComE+ [9] is another approach for community embedding which can tackle the situation in which the number of communities is unknown. However, these methods focus on node and community embeddings based on intra-group connections in terms of clustering and node classification tasks rather than link prediction.

Details of more relevant works are discussed in several surveys [6, 11]. A comprehensive review of recent approaches for knowledge graph representation is available in a survey paper [19], where KG embedding models are discussed in terms of representation space, scoring function, encoding models, and auxiliary information. The survey also discusses a broad range of reasoning methods such as Random Walk inference, Deep Reinforcement learning for multi-hop reasoning, and rule-based reasoning which however mainly focus on path-related reasoning rather than link prediction.

2.4. Scholarly Knowledge Graphs

In the last years, we saw the emergence of several knowledge graphs describing research publications.Traditionally, they either focus on the metadata of the articles, such as titles, abstracts, authors, organizations, or, more rarely, they offer a machine-readable representation of the knowledge contained therein.

A good example of the first category is Microsoft Academic Graph (MAG) [51], which is a heterogeneous knowledge graph containing the metadata of more than 242M scientific publications, including citations, authors, institutions, journals, conferences, and Fields of Study. Similarly, the Semantic Scholar Open Research Corpus[9] [1] is a dataset of about 185M publications released by Semantic Scholar, an academic search engine provided by the Allen Institute for Artificial Intelligence. The OpenCitations Corpus [37] is released by OpenCitations, an independent infrastructure organization for open scholarship dedicated to the publication of open bibliographic and citation data with semantic technologies. The current version includes 55M publications and 655M citations. Scopus is a well-known dataset curated by Elsevier, which includes about 70M publications and is often used by governments and funding bodies to compute performance metrics. The AMiner Graph [62] is a corpus of more than 200M publications generated and used by the AMiner system[10]. AMiner is a free online academic search and mining system that also extracts researchers' profiles from the Web and integrates them in the metadata. The Open Academic Graph (OAG)[11] is a large knowledge graph integrating Microsoft Academic Graph and AMiner Graph.

[9] ORC - http://s2-public-api-prod.us-west-2.elasticbeanstalk.com/corpus/

[10] AMiner - https://www.aminer.cn/

[11] OAG - https://www.openacademic.ai/oag/

The current version contains 208M papers from MAG and 172M from AMiner. Core [20][12] is a repository that integrates 24M open access research outputs from repositories and journals worldwide. The Dimensions Corpus is a dataset produced by Digital Science which integrates and interlinks 109M research publications, 5.3M grants, and 40M patents.

All these resources suffer from different degrees of data incompleteness. For instance, it is still challenging to identify and disambiguate affiliations, which also hinders the ability to categorize the articles according to their affiliation types or countries [27]. Similarly, references are usually incomplete, and the citation count of the same paper tends to vary dramatically on different datasets [37].

A second category of knowledge graphs focuses instead on representing the content of scientific publications. This challenging objective was traditionally pursued by the semantic web community, e.g., by creating bibliographic repositories in the Linked Data Cloud [33], generating knowledge bases of biological data [4], encouraging the Semantic Publishing paradigm [42], formalising research workflows [56], implementing systems for managing nanopublications [16, 22] and micropublications [41], developing a variety of ontologies to describe scholarly data, e.g., SWRC[13], BIBO[14], BiDO[15], SPAR [36][16], CSO[17] [40]. A recent example is the Open Research Knowledge Graph (ORKG) [18][18], which aims to describe research papers in a structured manner to make them easier to find and compare. Similarly, the Artificial Intelligence Knowledge Graph (AI-KG) [12][19] describes 1.2M statements extracted from 333K research publications in the field of AI. Since extracting the scientific knowledge from research articles is still a very challenging task, these resources tend also to suffer from data incompleteness. Therefore, it is crucial to develop new models that could tackle this issue and improve the quality of these KGs.

3. Motivating Scenario: AIDA Knowledge Graph

3.1. Incompleteness in AIDA

Academia, industry, public institutions, and non-profit organizations collaborate in the crucial effort of advancing scientific knowledge. Analysing the knowledge flow between them, assessing the best policies to harmonise their efforts, and detecting how they address emerging research areas is a critical task for researchers, funding bodies, and companies in the space of innovation. However, today scholarly KGs [37, 51, 1, 62, 20] do not support well this task since they typically lack a high-quality characterization of the research topics, affiliation types, and industrial sectors.

[12]CORE - https://core.ac.uk/
[13]SWRC - http://ontoware.org/swrc
[14]BIBO - http://bibliontology.com
[15]BiDO - http://purl.org/spar/bido
[16]SPAR - http://www.sparontologies.net/
[17]CSO - http://http://cso.kmi.open.ac.uk
[18]ORKG - https://www.orkg.org/orkg/
[19]AI-KG - http://scholkg.kmi.open.ac.uk/

Table 1
AIDA - Number of documents associated with the main GRID types.

	Papers	Patents
Education	3,969,097	169,884
Company	954,143	5,335,836
Government	185,633	54,396
Facility	169,234	66,605
Nonprofit	61,129	38,959
Healthcare	28,362	28180
Other	25,028	16,631
Typed (GRID)	5,133,171	5,639,252
Total	20,850,710	7,940,034

Therefore, we recently introduced the Academia/Industry DynAmics (AIDA) Knowledge Graph [2], which includes more than one billion triples and describes 20M publications from Microsoft Academic Graph (MAG)[20] [51] and 8M patents from Dimensions[21] according to the 14K research topics from the Computer Science Ontology (CSO)[22] [39]. In addition, 5.1M publications and 5.6M patents that were associated with IDs from the Global Research Identifier Database (GRID)[23] in the original data were also classified according to the type of the author's affiliations and 66 industrial sectors (e.g., automotive, financial, energy, electronics) drawn from the Industrial Sectors ontology (INDUSO)[24]. The mapping with MAG enables to characterize all articles according to the relevant scholarly entities in the MAG [51, 13], including *authors*, *conferences*, *journals*, *references* (the full citation network), and the *Fields of Study*. In the following, we will refer to the combination of the two knowledge graph as *AIDA+MAG*.

AIDA is available at `http://aida.kmi.open.ac.uk` and can be downloaded as a dump or queried via a Virtuoso triplestore (`http://aida.kmi.open.ac.uk/sparql/`). The AIDA ontology builds on SKOS, CSO, and INDUSO and it is available at `http://aida.kmi.open.ac.uk/ontology`.

Table 1 shows the number of publications and patents associated with the main categories from GRID. A document can be associated with multiples types according to the affiliations of the creators. Academic institutions ('education') are responsible for the majority of research publications (77.5%), while companies contribute to 19.8% of them. When considering patents, the picture is very different: 94.6% of them are from companies and only 3.0% are typed as 'education'.

AIDA was specifically designed for analysing the evolution of research topics across academia, industry, government facilities, and other institution. For instance, Figure 1 shows an example of the most frequent 16 high-level topics and reports the relevant percentage of academic publications, industry publications, academic patents, and indus-

[20]MAG - https://academic.microsoft.com/
[21]Dimensions - https://www.dimensions.ai/
[22]CSO - http://cso.kmi.open.ac.uk/
[23]GRID - https://www.grid.ac/
[24]INDUSO - http://aida.kmi.open.ac.uk/downloads/induso.ttl

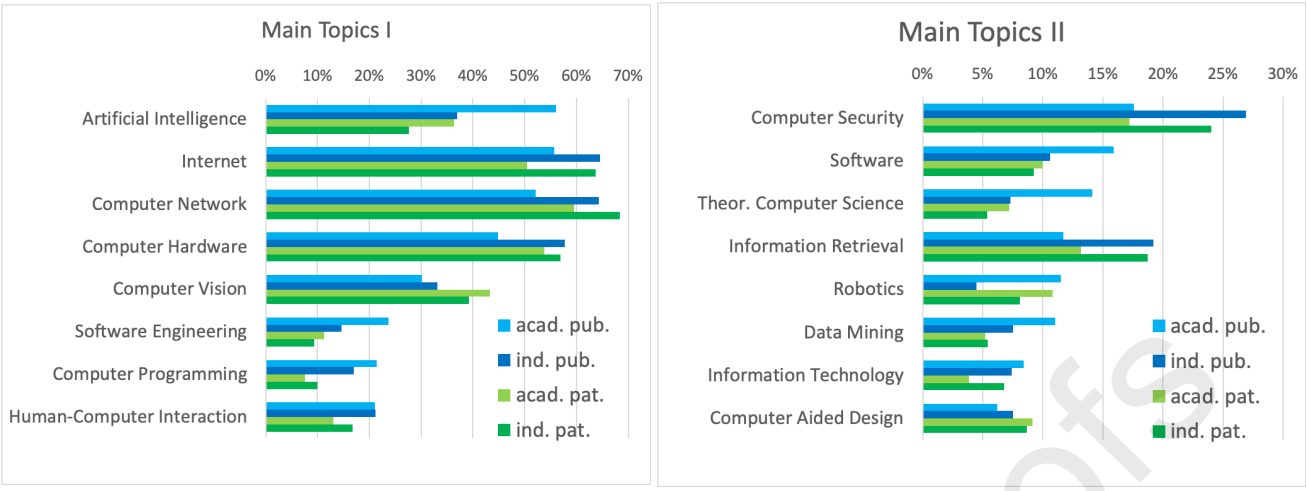


Figure 1: Distribution of the main topics in academia and industry.

trial patents. Some topics, such as Artificial Intelligence and Theoretical Computer Science, are mostly addressed by academic publications. Others, e.g., Computer Security, Computer Hardware, and Information Retrieval, attract a stronger interest from the industry. The topics which are mostly associated with patents are Computer Networks, Internet, and Computer Hardware. The overall sum of percentages for a given category on a certain topic may be more than 100% because each document may be associated with multiple topics.

The current version only associates the affiliation type with about 5.1M out of 21M articles and 5.6M out of the 8M patents. The missing documents that could not be typed either have affiliations that are not present on GRID, or they were not correctly mapped to the relevant GRID ids in the original data, for instance, because it was not possible to identify the institution when parsing the article. This is an exemplary case of the incompleteness problem which ultimately affects all scholarly knowledge graphs: a field that could enable crucial analyses exists only for about 25% of the documents.

This pragmatic scenario motivated us to investigate the best models for link prediction that could be applied on AIDA+MAG and by extensions on other knowledge graphs that suffer from similar issues. AIDA, as many KGs in the scholarly domain, describes millions of documents according to a relatively small set of categories, making this task quite challenging for two main reasons. First, there is an abundance of N to M relations with $N \gg M$, a situation that is not well handled by the current solutions, as we will further discuss in the next section. In addition, the number of items makes computationally unfeasible the adoption of embeddings with large dimensions. Therefore, it is critical to develop a solution that is able to both handle the $N \gg M$ issue and work well with limited dimensions. These considerations led to designing the novel Trans4E model presented in this paper.

Since we also wanted to test our solution on a field with different characteristics, we included in our analysis also the Fields of Study[25], which are terms from an in-house taxonomy used by MAG to index research papers. While we found that the topics from CSO typically produce a better representation of the domain of Computer Science [39] and thus are more apt for the analyses that focus on this discipline, FoS has the advantage of covering all research fields. In the context of AIDA this is particularly useful for characterizing industrial sectors and companies that are interested in Computer Science applications in other fields such as Medicine, Chemistry, Geology, and Physics. Furthermore, FoS is a very interesting case study for a variety of reasons. First, its purpose is the description of the knowledge in the articles rather than their objective characteristics, such as the list of authors or the venues. Second, it is already present in the vanilla MAG, allowing us to experiment on a well established KG in this space. Finally, the quality of the topics for a specific document depends on the length and style of the abstract, which is typically parsed by an online PDF file, sometimes leading to mistakes and missing content. Indeed, documents that are associated with short or incorrectly parsed abstracts are usually tagged with very few topics. Considering the relevant entities in the graph such as authors, venues, and references, may enable to identify the missing topics.

Completing the Fields of Study is critical for supporting deeper analyses of the trends of multiple disciplines. Besides, the same approach may be adopted to support the integration of documents from other knowledge graphs that lack abstracts (e.g., DBLP[26], OpenCitations[27]). For the sake of simplicity, in the rest of the paper we will refer to the Fields of Study from MAG in AIDA+MAG simply as *topics*.

3.2. Limitations of KGEs for Link Prediction

To improve the number of documents in AIDA+MAG characterized according to their affiliation types, we tried four well-know embedding models: TransE, RotatE, ComplEx and QuatE. Their performance on this KG was not par-

[25] Fields of Study - https://academic.microsoft.com/topics/
[26] DBLP - https://dblp.org/
[27] OpenCitations - https://opencitations.net/

ticularly good, in particularly when using small embedding vectors.

A systematic analysis showed that this was due to the characteristics of the *hasGRIDType* relation: a N to M relation where N is large and M small, as in M = 7: 'Education', 'Government', 'Company', 'Healthcare', 'Facility','Nonprofit', 'Other'. Indeed, for any triple in the form $(h, hasGRIDType, t)$, there is a large number N of h entities (papers) given a specific tail t (types), and a small number M of t (types) given a specific head h (papers). The same applies for the triples with the *hasTopic* relation that associates articles with topics.

The TransE model has some issues when handling these kinds of relations. Let us consider N entities h (papers) typed as 'Education' $(h, hasGRIDType, Education)$. According to the TransE formulation 1 (see Equation 1), this case can be formalized as:

$$\begin{cases} \mathbf{h}_1 + \mathbf{r} = \mathbf{t}, \\ \mathbf{h}_2 + \mathbf{r} = \mathbf{t}, \\ \vdots \\ \mathbf{h}_N + \mathbf{r} = \mathbf{t}. \end{cases} \tag{7}$$

Since the relation $\mathbf{r}$ is always *hasGRIDType* and the tail $\mathbf{t}$ always *Education*, the formulation of the model enforces to have $\mathbf{h}_1 = \mathbf{h}_2 = \ldots = \mathbf{h}_N$ for N number of papers. This is an issue since the embedding vectors of all the entities appearing in the head will be very similar. Consequently, the model may be unable to distinguish among them, resulting in poor performance. Therefore, the TransE model is more suitable for N to M relations in which N and M are both small.

RotatE suffers from the same issue. According to the RotatE formulation, we have:

$$\begin{cases} \mathbf{h}_{1i} \circ \mathbf{r}_i = \mathbf{t}_i, \\ \mathbf{h}_{2i} \circ \mathbf{r}_i = \mathbf{t}_i, \\ \vdots \\ \mathbf{h}_{Ni} \circ \mathbf{r}_i = \mathbf{t}_i, \qquad i = 1, \ldots, d. \end{cases} \tag{8}$$

This shows that in each element of the embedding vector (indexed as i), there is only one option for embeddings for any head for a given tail. Therefore, in the case of the *hasGRIDType* relation (also for *hasTopic*), RotatE lacks the capacity to distinguish well among the research papers in the head (h). In conclusion, a larger vector space appears to be crucial to properly represent these kinds of relations and perform high-quality link prediction on AIDA and similar scholarly knowledge graphs.

Other models, such as ComplEx and QuatE, suffer from two major issues when applied on KG with high number of entities: a) since they use 1 to K negative sampling [23] (where K is the total number of entities in a KG), in the case of N to M relations where N $\gg$ M, a substantial portion of the samples are positive but they are used as negative samples; b) high computation costs also resulted from using K negative sampling.

4. Methodology

4.1. The Trans4E Model

Trans4E is a novel KGE model designed to effectively handle KGs which include N to M relations with N$\gg$M. In this section, we show that the capacity of this model for a given relation (e.g., *hasGRIDType*, *hasTopic*) and the corresponding tail entity (e.g., *type* or *topic*) is 8^d, which allows to generate a distinct vector for each entity (e.g., a specific paper) even when using small embedding dimensions. Here we introduce the core formulation of the score function of Trans4E.

Trans4E maps the entities of the graph via relations in Quaternion vector space $\mathbb{H}^d$. Concretely, given a triple of the form (h, r, t), our model follows the following steps:

(a) The head entity vector ($\mathbf{h} \in \mathbb{H}^d$) is rotated by $\mathbf{r}_\theta$ degrees in quaternion space i.e. $\mathbf{h}_{\theta_r} = \mathbf{h} \otimes \mathbf{r}_\theta$. $\otimes$ is an element-wise Hamilton product between two quaternion vectors.

(b) The rotated head i.e. $\mathbf{h}_{\theta_r}$ is translated by the relation embedding vector $\mathbf{r}$ to get $\mathbf{h}_r = \mathbf{h}_{\theta_r} + \mathbf{r}$.

(c) The translated head embedding vector should meet the tail embedding vector i.e. $\mathbf{h}_r \approx \eta_h \otimes \mathbf{t}$ for a positive sample (h, r, t). $\mathbf{t} \in \mathbb{H}^d$. However, there is a possibility that the transformed vector of the head is not exactly meeting the tail. In order to solve this problem, we could use $\eta_h = [\eta_{h1}, \ldots, \eta_{hd}] \in \mathbb{H}^d$, which is a mapping regularizer.

Following the mentioned steps, we define the score function as:

$$f(\mathbf{h}, \mathbf{r}, \mathbf{t}) = -\|\mathbf{h}_r - \eta_h \otimes \mathbf{t}\|. \tag{9}$$

The score function returns a low value if the triple is false i.e. $\mathbf{h}_r \neq \eta_h \otimes \mathbf{t}$ and returns high value (close to zero) if the triple is true i.e. $\mathbf{h}_r \approx \eta_h \otimes \mathbf{t}$. In this way, we measure the plausibility of each triple (h, r, t).

In addition, two regularized versions of the Trans4E model are also made available. The first is Trans4EReg1, which is the regularized version of Trans4E where a relation-specific head rotation and the tail mapping regularizer are used. The second is Trans4EReg2, which is a regularized version of Trans4E with a relation-specific rotation on the tail side (in addition to the relation-specific head rotation and the tail mapping regularizer).

4.2. Link Prediction on N to M Relations

Here we show that Trans4E provides a higher capacity with fewer limitations than other models. Given a relation r (e.g., *hasGRIDType*) and a tail t (e.g., 'Education'), the following constraints are applied for each of the resulting triples:

$$\begin{cases} \mathbf{h}_{1\theta_{ri}} + \mathbf{r}_i = \eta_{h_{1i}} \otimes \mathbf{t}_i, \\ \mathbf{h}_{2\theta_{ri}} + \mathbf{r}_i = \eta_{h_{2i}} \otimes \mathbf{t}_i, \\ \vdots \\ \mathbf{h}_{N\theta_{ri}} + \mathbf{r}_i = \eta_{h_{Ni}} \otimes \mathbf{t}_i, \qquad i = 1, \dots, d. \end{cases} \tag{10}$$

We can rewrite the Hamilton product as 4-dimensional matrix-vector product:

$$\mathbf{h}_{\theta_{ri}} = \mathbf{h}_i \otimes \mathbf{r}_{\theta_i} = \begin{bmatrix} a_{r_\theta} & -b_{r_\theta} & -c_{r_\theta} & -d_{r_\theta} \\ b_{r_\theta} & a_{r_\theta} & -d_{r_\theta} & c_{r_\theta} \\ c_{r_\theta} & d_{r_\theta} & a_{r_\theta} & -b_{r_\theta} \\ -d_{r_\theta} & -c_{r_\theta} & b_{r_\theta} & a_{r_\theta} \end{bmatrix} \begin{bmatrix} a_h \\ b_h \\ c_h \\ d_h \end{bmatrix} = \mathcal{H}_i \vec{h}_i. \tag{11}$$

Without loss of generality, we assume that the embedding of the relation translation $\mathbf{r}_i$ is zero and $\eta_{h_{pi}}$ is a real value. In this way, we can write the above system of equations in the following form:

$$\begin{cases} \mathcal{H}_i \vec{h}_{1i} = \eta_{h_{1i}} \vec{t}_i \\ \mathcal{H}_i \vec{h}_{2i} = \eta_{h_{2i}} \vec{t}_i \\ \vdots \\ \mathcal{H}_i \vec{h}_{1i} = \eta_{h_{Ni}} \vec{t}_i, \qquad i = 1, \dots, d. \end{cases} \tag{12}$$

Note that the matrix $\mathcal{H}_i$ is 4×4 and has 4 distinct eigenvalues/eigenvectors. Therefore, we can write $\mathcal{H}_i \vec{h}_{pi} = \lambda_{h_{pi}} \vec{h}_{pi} = \eta_{h_{pi}} \vec{t}_{pi}$. If $\lambda_{h_{pi}} = \eta_{h_{pi}}$, then the ith dimension of the head and tail vectors will be same, otherwise, they will be different. Therefore, we will have $4\,(\text{number of distinct eigenvectors}) \times 2 = 8$ various options in each dimension to be assigned to the head entity vector. The multiplication by 2 is due to the two possible cases, one for the equality of the head and the tail and the other for their inequality.

Because we use d dimensional vectors, we have $8^d - 1$ possible distinct vectors to be assigned to the entities appearing in the head (e.g., articles in AIDA). As a result, the capacity of the model becomes $8^d - 1$, which provides a larger space than the TransE and RotatE models. In Section 5, we will show the advantages of this solution by comparing it against alternative models.

5. Evaluation

We compared Trans4E against four alternative embedding models: TransE, RotatE, ComplEX, and QuatE.

5.1. Evaluation Datasets

We ran the experiments on a portion of the knowledge graph *AIDA+MAG* including 68,906 entities and 180K triples. Specifically, we considered the following entities: publication IDs, authors, affiliation organizations, topics, publication types, conference editions, conference series, journals, years, countries, and references.

In this subset, the *hasGRIDType* relation includes about 5k entities (research papers) in the head position and 7 entities as tail ('Education', 'Company', 'Government', 'Healthcare', 'Nonprofit', 'Facility', and 'Other'). Regarding the *hasTopic* relation, the highest number of research articles associated to a topic is 4,659, while the highest number of topics associated to research articles is only 13.

We split the datasets into train (80%), test (10%), and validation (10%) sets. Additionally, we evaluated the performance of our model on four benchmarks: FB15K (14,951 entities and 1,345 relations), FB15k-237 (14,451 entities and 237 relations), WN18 (40,943 entities and 18 relations), and WN18RR (40,943 entities and 13 relations).

5.2. Evaluation Criteria

In this section we discuss the criteria that we considered for the evaluation.

Performance Metrics. The standard evaluating metrics for the performance of KGEs are: Mean Rank (MR), Mean Reciprocal Rank (MRR) and Hits@k (k=1, 3, 10) [52]. MR is the average rank of correct triples in the test set. In order to compute it, we generate two sets of triples, $S_h = (h, r, ?)$ and $S_t = (?, r, t)$, by corrupting each test triple (h, r, t). After this step, the scores of all the triples in S_h, S_t are computed and the triples are sorted. The rank (r_h, r_t) of the original triple (i.e. (h, r, t)) is then computed in both sets S_h, and S_t. For any triple, r_h is the notation for the right ranks and r_t for the left ranks. The rank of the example triple of (h, r, t) is computed as $rank = \frac{r_h + r_t}{2}$. If we assume $rank_i$ to be the rank of the $i-$th triple in the test set obtained by a KGE model, then the MR and the MRR are obtained as follows:

$$MR = \sum_i rank_i,$$

$$MRR = \sum_i \frac{1}{rank_i}.$$

For the evaluation on *hasGRIDType* and *hasTopic* relations, we only corrupted the tail of the relations and replaced it with all the entities in the KG.

The $Hits@K$, for k = 1, 3, 10 ..., is one of the standard link prediction measurements. By considering the percentage of the triples for which $rank_i$ is equal or smaller than k, we computed the $Hits@K$. MR, the average MRR, Hits@1, Hits@3, and Hits@10 are reported in Tables 2-6.

Dimension and KG Scale. Although the performance measures of a machine learning model are important criteria for evaluation, the dimension of the embedding vectors is specifically important for KGE models, which are supposed to be used in the real-world large-scale KGs. Indeed, an embedding with very large dimensions may be unfeasible in most practical settings.

Therefore, we compared the performances of our model against state-of-the art models in a very low dimensional embedding. This was done to simulate a real-world application

Model Type	hasTopic					hasGRIDType				
	MR	MRR	Hits@1	Hits@3	Hits@10	MR	MRR	Hits@1	Hits@3	Hits@10
TransE	3785	0.031	0.006	0.027	0.071	6	0.658	0.500	0.771	0.970
RotatE	4749	0.036	0.000	0.001	0.008	38	0.472	0.000	0.000	0.001
QuatE	4862	0.066	0.021	0.066	0.151	159	0.252	0.166	0.271	0.431
ComplEx	3726	0.044	0.003	0.042	0.111	6	0.429	0.001	0.838	0.931
Trans4EReg1	3007	0.403	0.325	0.450	0.531	1	0.941	0.915	0.978	0.995
Trans4EReg2	2047	0.401	0.325	0.445	0.528	1	0.956	0.928	0.985	0.988
Trans4E	2908	0.089	0.030	0.083	0.211	1	0.900	0.834	0.965	0.998

Table 2
Performance of KGEs on AIDA for Dimension 5

Model Type	hasTopic					hasGRIDType				
	MR	MRR	Hits@1	Hits@3	Hits@10	MR	MRR	Hits@1	Hits@3	Hits@10
TransE	3903	0.135	0.043	0.126	0.355	1	0.859	0.769	0.944	1.000
RotatE	3890	0.155	0.057	0.144	0.411	1	0.891	0.823	0.970	1.000
QuatE	1693	0.093	0.057	0.106	0.165	1718	0.096	0.062	0.116	0.148
ComplEx	7279	0.081	0.036	0.093	0.167	700	0.896	0.869	0.919	0.939
Trans4EReg1	2424	0.379	0.300	0.416	0.515	117	0.907	0.856	0.947	0.991
Trans4EReg2	3250	0.394	0.327	0.429	0.507	1	0.959	0.928	0.990	1.000
Trans4E	3842	0.158	0.053	0.154	0.416	1	0.866	0.790	0.931	1.000

Table 3
Performance of KGEs on AIDA for Dimension 50

of KGEs on large scale KGs. Indeed, models which obtain satisfactory performances on a portion of a graph using a small vector size should also perform well when adopting a higher dimension on a larger portion of the same graph [30, 15].

5.3. Hyperparameter Setting

The development environment of our model is PyTorch[28]. In the experiments, we reshuffled the training set in each epoch, and generated 16 mini batches on the reshuffled samples. To determine the performances of our model in high and low dimensions, the embedding dimension (d) was set to $\{5, 50, 500\}$ in the experiments. The batch size (b) is considered as $\{256, 512\}$, the fixed margin γ is $\{2, 3, 4, 5, 10, 15, 20, 30\}$ and learning rate as $\{0.001, 0.01, 0.05, 0.1\}$ with a negative sample of 10. L_2 regularization coefficient is $\{0.000005, 0.0000005\}$ for the models QuatE, Trans4EReg1, and Trans4EReg2. The best hyperparameter combination for Trans4E and Trans4EReg2 is $b = 256$, $lr = 0.1$, $\gamma = 20$ and for Trans4EReg1 is $b = 256$, $lr = 0.001$, $\gamma = 20$, and $d = 500$ for all the models. For the regularized versions λ =0.000005.

[28] PyTorch - https://pytorch.org/

5.4. Results and Discussions

In this section, we present the results of our experiments. Specifically, Section 5.4.1 reports the results of the evaluation regarding the graph completion on AIDA+MAG. Section 5.4.2 compares the performance of Trans4E and several alternatives on a set of four standard benchmarks (FB15k, FB15k-237, WN18, andWN18RR). Section 5.4.3 investigates the representation of the research topics and shows a

Model Type	hasTopic					hasGRIDType				
	MR	MRR	Hits@1	Hits@3	Hits@10	MR	MRR	Hits@1	Hits@3	Hits@10
TransE	3982	0.400	0.294	0.462	0.592	1	0.968	0.944	0.990	1.000
RotatE	4407	0.433	0.332	0.492	0.622	1	0.953	0.933	0.975	0.996
QuatE	1353	0.426	0.341	0.472	0.581	1	0.957	0.928	0.983	0.998
ComplEx	5855	0.099	0.077	0.109	0.129	1566	0.566	0.531	0.596	0.609
Trans4EReg1	2040	0.402	0.295	0.466	0.604	233	0.910	0.882	0.937	0.944
Trans4EReg2	1942	0.424	0.325	0.482	0.602	34	0.955	0.931	0.978	0.990
Trans4E	3904	0.426	0.318	0.492	0.628	1	0.968	0.944	0.995	0.998

Table 4
Performance of KGEs on AIDA for dimension 500.

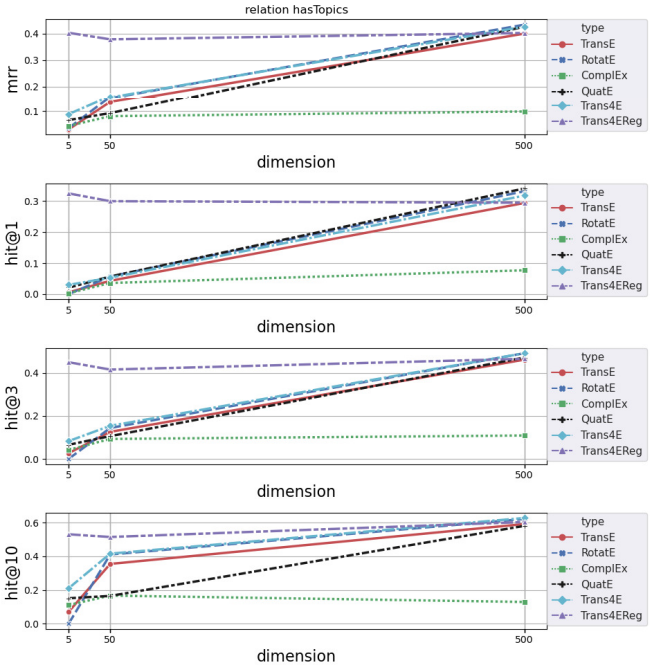


Figure 2: hasTopic for dimension 5,50 and 500

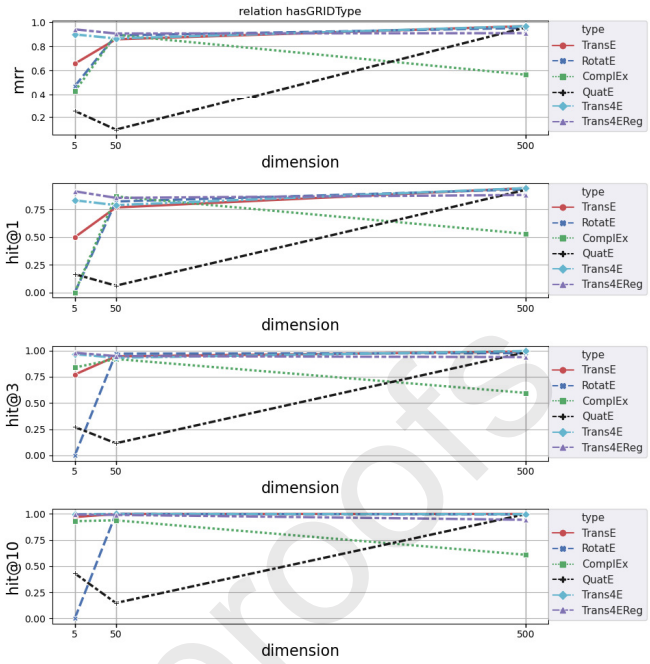


Figure 3: hasGRIDType for dimension 5,50 and 500

study of the distribution of their embedding vectors.

5.4.1. Knowledge Graph Completion in AIDA+MAG

In this section we evaluate the performance of Trans4E versus alternative methods in completing the two relations *hasGRIDType* and *hasTopic* in AIDA+MAG. Specifically, we compared Trans4E with TransE, RotatE, QuatE and ComplEx. We also included Trans4EReg1 and Trans4EReg2, the two regularized versions previously defined in Section 4.1.

Table 2 reports the performances of the seven models for dimension 5. Trans4EReg1 clearly outperforms all the other models for the *hasTopic* relations. Trans4EReg2 obtains the second-best performance. For instance, when considering the *hasTopic* relation, Trans4EReg1 and Trans4EReg2 yield 32.5% in Hits@1 while all the other solutions obtain less than 3%. For the *hasGRIDType* relations Trans4EReg2 outperforms all the others with a 92.8% in Hits@1. Moreover, Trans4EReg1, yields 91.5% in Hits@1 and Trans4E 83.4%, while the best of the other models is TransE with 50.0%.

RotatE performed surprisingly poorly on both the *hasTopic* and *hasGRIDType* relations, yielding 0% in Hits@1. It should be noted that during testing, for a test triple $(p, hasGRIDType, t)$, we replaced the tail t with all the entities in the graph and ranked the actual ones against the corrupted triples. As a result, RotatE (in dimension 5) does not rank any type entity even among the top 10 occurrences. This means that non-type entities in the corruption process are ranked higher than the typed entities. This is related to the limited solution space of the RotatE model, which is also discussed in [31].

The overall accuracy for *hasGRIDType* is typically higher than *hasTopic*. For instance, Trans4EReg1 yields a Hits@10 of 99.5% for *hasGRIDType* and 53.1% for *hasTopic*. This is mainly due to the fact that the number of entities to be considered for *hasTopic* is much higher than that for *hasGRIDType*.

Overall, Trans4EReg1 seems to be the most suitable model for addressing large-scale KGs, where increasing the dimension of the model is too costly in computational terms.

Table 3 reports the performances of the models using dimensions 50. Trans4EReg1 and Trans4EReg2 outperforms all the models with regards to the *hasTopic* by a considerable margin (up to 10% improvement on Hits@10). When considering *hasGRIDType*, Trans4EReg2 obtains the best performances in all metrics, folowed by Trans4EReg1 and RotatE. Due to the overfitting, the performance of Trans4EReg1 and Trans4EReg2 decreases as the dimension increases from 5 to 50. In fact, Trans4EReg1 and Trans4EReg2 with dimension 5 still outperforms all the models with dimension 50 in most of the metrics.

Table 4 reports the experiments with a dimension of 500. For *hasGRIDType*, Trans4E and TransE are comparable and obtain the best performances. When considering *hasTopic*, QuatE, RotatE, and Trans4E perform similarly well. Specifically, QuatE yields the best performance in Hits@1 (34.1%), while Trans4E and RotatE perform best in Hits@3 (49.2%), and Trans4E obtains the highest Hits@10 (62.8%).

Figure 2 and 3 summarize the performances of all the models for dimension 5, 50, and 500. Trans4EReg1 significantly outperforms all the models when using low dimensions and performs well also in high dimensions.

Model Type		FB15k				WN18				
	MR	MRR	Hits@1	Hits@3	Hits@10	MR	MRR	Hits@1	Hits@3	Hits@10
TransE	–	0.463	0.297	0.578	0.749	–	0.495	0.113	0.888	0.943
RotatE	40	0.797	0.746	0.830	0.884	309	0.949	0.944	0.952	0.959
QuatE	35	0.742	0.658	0.805	0.881	349	0.942	0.927	0.952	0.960
Trans4E	47	0.767	0.681	0.834	0.892	175	0.950	0.944	0.953	0.960

Table 5
Performance of KGEs on FB15K and WN18.

Model Type		FB15k-237				WN18RR				
	MR	MRR	Hits@1	Hits@3	Hits@10	MR	MRR	Hits@1	Hits@3	Hits@10
TransE	357	0.294	–	–	0.465	3384	0.226	–	–	0.501
RotatE	177	0.338	0.241	0.375	0.533	3340	0.476	0.428	0.492	0.571
QuatE	170	0.282	0.178	0.315	0.501	2272	0.303	0.179	0.386	0.530
Trans4E	158	0.332	0.236	0.366	0.527	1755	0.469	0.416	0.487	0.577

Table 6
Performance of KGEs on FB15K-237 and WN18RR.

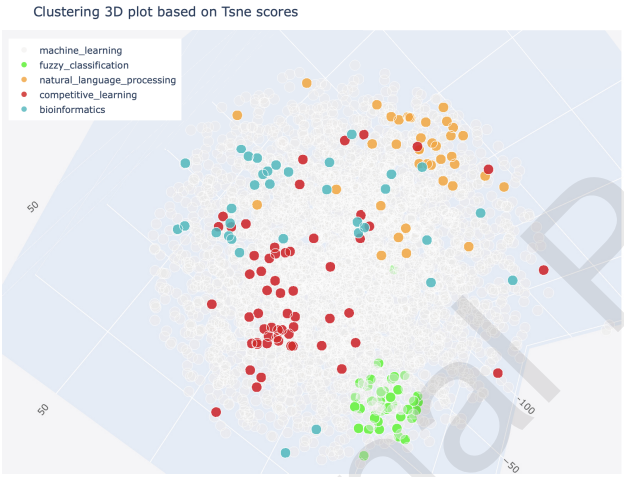


Figure 4: Distribution of the main topics in academia and industry.

5.4.2. *Link Prediction on Benchmark Datasets*

We evaluated the performances of the Trans4E model against the competitors on a set of standard benchmark datasets with diverse relations (N to M relations where N and M are large, N and M are small, N≫M and N≪M).

Table 5 and Table 6 show the performances of the KGE models on the benchmark datasets FB15k, FB15k-237, WN18, and WN18RR. Trans4E outperforms the other models in Hits@3 and Hits@10 in FB15k and WN18. It also obtains a significantly better MR on FB15k-237 and WN18RR. In FB15k, the Trans4E model outperforms all the other models when considering the Hits@3 and Hits@10. In WN18, Trans4E outperforms TransE and QuatE, and obtains competitive results with respect to RotatE. To note that, these results are computed by running the models on the benchmark datasets using the best obtained hyperparameter settings where the dimension is 200, and with 20 negative samples using adversarial negative sampling [45]. The results are comparatively close in the case of FB15k-237 and WN18RR, where Trans4E has a better performance in MR. Overall, the results show that our model outperforms other KGE models on N to M relations with N≫M and provides competitive performance on KGs with diverse relations.

5.4.3. *Efficiency of the Embeddings*

To further investigate the representation of research topics with Trans4E, we analysed how the embeddings discriminate articles tagged with different topics.

Figure 4 shows the embeddings associated to the articles in AIDA+MAG in two dimensions. In order to produce it, we first selected five major topics of the machine learning venues: "fuzzy_classification", "natural_language_processing", "competitive_learning", "machine learning", and "bioinformatics". Then, we retrieved the embedding vectors of the papers tagged with those topics and visualized them by using T-SNE [26].

We can appreciate how papers with the same topics tend to cluster together. For example, papers belonging to the "fuzzy_classification" topic (green) lie within the same cluster. Note that papers in some topics such as "bioinformatics" may be associated to other topics as well (e.g. a paper may be in "bioinformatics" and use "fuzzy_classification" methods). This is why papers related to more general topics are distributed with a larger variance.

We further evaluated the ability of our model to properly distribute topics in the vector space based on their publication dates. In Figure 5, we illustrate the distribution of the learned vectors for the topics w.r.t their publishing years. This shows that topics such as "convolutional_neural_networks", "parallel_processing", and "speech_recognition" are correctly identified to be hot topics

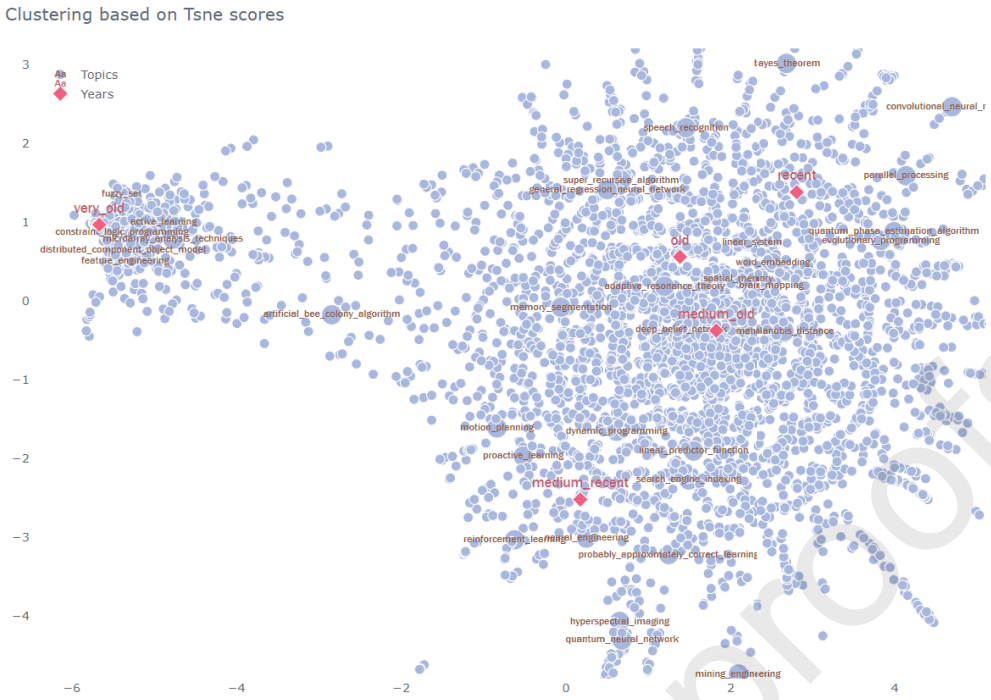


Figure 5: Distribution of Topics w.r.t Years. year >= 2015 is considered recent, year >= 2010 and year < 2015 are denoted as medium_recent, year >= 2005 and year < 2010 are medium_old, year>= 2000 and year < 2005 mean old, and anything before 2000 is very_old.

for the corresponding years.

The topic "word_embedding" lies in the border of recent and medium_old period indicating that even if old is still lasting.There is also a cluster of topics around the very_old time period for which the corresponding vectors are very different from the ones in other time periods. A manual analysis revealed that most of them were mostly active before the year 2000.

6. Conclusion and Future Work

In this paper we presented Trans4E, a KGE model designed to provide link prediction for KGs that include N to M relations with N≫M. Trans4E and its regularized versions (Trans4EReg1 and Trans4EReg2) have been applied on a real world case involving the Academic/Industry DynAmics Knowledge Graphs (AIDA) and the Microsoft Academic Graph (MAG). The evaluation showed that Trans4E outperforms other approaches in the case of N to M relations with N ≫ M and obtains competitive results in all the other settings, in particular when using low embedding dimensions. Hence, our approach appears to be an effective and generalizable solution able to achieve and sometimes improve state-of-the-art performance on many established benchmarks. In addition, it seems to be the most effective model when dealing with shallow classification schemes and using low embedding dimensions.

In future work we aim to perform link prediction on other relations in AIDA and MAG, with the aim to release a new version of these KGs for the research community. We also intend to use Trans4E for supporting a variety of other tasks involving scholarly KGs, such as trend detection, expert search, and recommendation of articles and patents. Finally, we plan to investigate the application of Trans4E on KGs describing scientific knowledge, such as AI-KG [12] and ORKG [18].

Acknowledgements

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X GPU used for this research. This work is supported by the EC Horizon 2020 grant LAMBDA (GA no. 809965), and the CLEOPATRA project (GA no. 812997).

References

[1] Ammar, W., Groeneveld, D., Bhagavatula, C., Beltagy, I., Crawford, M., Downey, D., Dunkelberger, J., Elgohary, A., Feldman, S., Ha, V., et al., 2018. Construction of the literature graph in semantic scholar. arXiv preprint arXiv:1805.02262 .

[2] Angioni, S., Salatino, A.A., Osborne, F., Recupero, D.R., Motta, E., 2020. Integrating knowledge graphs for analysing academia and industry dynamics, in: ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium, Springer. pp. 219–225.

[3] Arora, S., 2020. A survey on graph neural networks for knowledge graph completion. arXiv preprint arXiv:2007.12374 .

[4] Belleau, F., Nolin, M.A., Tourigny, N., Rigault, P., Morissette, J., 2008. Bio2rdf: towards a mashup to build bioinformatics knowledge systems. Journal of biomedical informatics 41, 706–716.

[5] Bendre, N., Marín, H.T., Najafirad, P., 2020. Learning from few samples: A survey. arXiv preprint arXiv:2007.15484 .

[6] Bonatti, P.A., Decker, S., Polleres, A., Presutti, V., 2019. Knowledge graphs: New directions for knowledge representation on the semantic web (dagstuhl seminar 18371), in: Dagstuhl Reports, Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

[7] Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., Yakhnenko, O., 2013. Translating embeddings for modeling multi-relational data, in: Advances in NIPS.

[8] Bose, A.J., Jain, A., Molino, P., Hamilton, W.L., 2019. Meta-graph: Few shot link prediction via meta learning. arXiv preprint arXiv:1912.09867 .

[9] Cavallari, S., Cambria, E., Cai, H., Chang, K.C.C., Zheng, V.W., 2019. Embedding both finite and infinite communities on graphs [application notes]. IEEE Computational Intelligence Magazine 14, 39–50.

[10] Chen, M., Zhang, W., Zhang, W., Chen, Q., Chen, H., 2019. Meta relational learning for few-shot link prediction in knowledge graphs, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 4208–4217.

[11] Chen, X., Jia, S., Xiang, Y., 2020. A review: Knowledge reasoning over knowledge graph. Expert Systems with Applications 141, 112948.

[12] Dessı, D., Osborne, F., Recupero, D.R., Buscaldi, D., Motta, E., Sack, H., . Ai-kg: an automatically generated knowledge graph of artificial intelligence .

[13] Färber, M., 2019. The microsoft academic knowledge graph: A linked data source with 8 billion triples of scholarly data, in: International Semantic Web Conference, Springer. pp. 113–129.

[14] Gong, K., Gao, Y., Liang, X., Shen, X., Wang, M., Lin, L., 2019. Graphonomy: Universal human parsing via graph transfer learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7450–7459.

[15] Goyal, P., Ferrara, E., 2018. Graph embedding techniques, applications, and performance: A survey. Knowledge-Based Systems 151, 78–94.

[16] Groth, P., Gibson, A., Velterop, J., 2010. The anatomy of a nanopublication. Information Services & Use 30, 51–56.

[17] Henk, V., Vahdati, S., Nayyeri, M., Ali, M., Yazdi, H.S., Lehmann, J., 2019. Metaresearch recommendations using knowledge graph embeddings, in: RecNLP workshop of AAAI Conference.

[18] Jaradeh, M.Y., Oelen, A., Farfar, K.E., Prinz, M., D'Souza, J., Kismihók, G., Stocker, M., Auer, S., 2019. Open research knowledge graph: Next generation infrastructure for semantic scholarly knowledge, in: Proceedings of the 10th International Conference on Knowledge Capture, pp. 243–246.

[19] Ji, S., Pan, S., Cambria, E., Marttinen, P., Yu, P.S., 2020. A survey on knowledge graphs: Representation, acquisition and applications. arXiv preprint arXiv:2002.00388 .

[20] Knoth, P., Zdrahal, Z., 2011. Core: connecting repositories in the open access domain, in: CERN Workshop on Innovations in Scholarly Communication (OAI7). URL: http://oro.open.ac.uk/32560/. poster Session ID: 53.

[21] Knoth, P., Zdrahal, Z., 2012. Core: three access levels to underpin open access. D-Lib Magazine 18, 1–13.

[22] Kuhn, T., Chichester, C., Krauthammer, M., Queralt-Rosinach, N., Verborgh, R., Giannakopoulos, G., Ngomo, A.C.N., Viglianti, R., Dumontier, M., 2016. Decentralized provenance-aware publishing with nanopublications. PeerJ Computer Science 2, e78.

[23] Lacroix, T., Usunier, N., Obozinski, G., 2018. Canonical tensor decomposition for knowledge base completion, in: International Conference on Machine Learning, pp. 2863–2872.

[24] Lee, J., Kim, H., Lee, J., Yoon, S., 2017. Transfer learning for deep learning on graph-structured data., in: AAAI, pp. 2154–2160.

[25] Li, L., Wang, P., Wang, Y., Jiang, J., Tang, B., Yan, J., Wang, S., Liu, Y., 2019. Prtransh: Embedding probabilistic medical knowledge from real world emr data. arXiv preprint arXiv:1909.00672 .

[26] Maaten, L.v.d., Hinton, G., 2008. Visualizing data using t-sne. Journal of machine learning research 9, 2579–2605.

[27] Mannocci, A., Osborne, F., Motta, E., 2019. Geographical trends in academic conferences: An analysis of authors' affiliations. Data Science 2, 181–203.

[28] Nathani, D., Chauhan, J., Sharma, C., Kaul, M., 2019. Learning attention-based embeddings for relation prediction in knowledge graphs, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 4710–4723.

[29] Nayyeri, M., Vahdati, S., Zhou, X., Yazdi, H.S., Lehmann, J., 2020a. Embedding-based recommendations on scholarly knowledge graphs, in: European Semantic Web Conference, Springer. pp. 255–270.

[30] Nayyeri, M., Xu, C., Vahdati, S., Vassilyeva, N., Sallinger, E., Yazdi, H.S., Lehmann, J., 2020b. Fantastic knowledge graph embeddings and how to find the right space for them, in: International Semantic Web Conference, Springer. pp. 438–455.

[31] Nayyeri, M., Xu, C., Vahdati, S., Vassilyeva, N., Sallinger, E., Yazdi, H.S., Lehmann, J., 2020c. Fantastic knowledge graph embeddings and how to find the right space for them, in: ISWC.

[32] Nickel, M., Murphy, K., Tresp, V., Gabrilovich, E., 2016. A review of relational machine learning for knowledge graphs. Proceedings of the IEEE 104.

[33] Nuzzolese, A.G., Gentile, A.L., Presutti, V., Gangemi, A., 2016. Semantic web conference ontology-a refactoring solution, in: European Semantic Web Conference, Springer. pp. 84–87.

[34] Osborne, F., Mannocci, A., Motta, E., 2017. Forecasting the spreading of technologies in research communities, in: Proceedings of the Knowledge Capture Conference, ACM, New York, NY, USA. pp. 1:1–1:8. doi:10.1145/3148011.3148030.

[35] Paliwal, A., Gimeno, F., Nair, V., Li, Y., Lubin, M., Kohli, P., Vinyals, O., 2019. Regal: Transfer learning for fast optimization of computation graphs. arXiv preprint arXiv:1905.02494 .

[36] Peroni, S., Shotton, D., 2018. The spar ontologies, in: International Semantic Web Conference, Springer. pp. 119–136.

[37] Peroni, S., Shotton, D., 2020. Opencitations, an infrastructure organization for open scholarship. Quantitative Science Studies 1, 428–444.

[38] Salatino, A., Osborne, F., Motta, E., 2020a. Researchflow: Understanding the knowledge flow between academia and industry, in: Knowledge Engineering and Knowledge Management – 22nd International Conference, EKAW 2020.

[39] Salatino, A.A., Thanapalasingam, T., Mannocci, A., Birukou, A., Osborne, F., Motta, E., 2020b. The computer science ontology: A comprehensive automatically-generated taxonomy of research areas. Data Intelligence 2, 379–416.

[40] Salatino, A.A., Thanapalasingam, T., Mannocci, A., Osborne, F., Motta, E., 2018. The computer science ontology: a large-scale taxonomy of research areas, in: ISWC, pp. 187–205.

[41] Schneider, J., Ciccarese, P., Clark, T., Boyce, R.D., 2014. Using the micropublications ontology and the open annotation data model to represent evidence within a drug-drug interaction knowledge base.

[42] Shotton, D., 2009. Semantic publishing: the coming revolution in scientific journal publishing. Learned Publishing 22, 85–94.

[43] Stanovsky, G., Gruhl, D., Mendes, P., 2017. Recognizing mentions of adverse drug reaction in social media using knowledge-infused recurrent models, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pp. 142–151.

[44] Sun, F.Y., Qu, M., Hoffmann, J., Huang, C.W., Tang, J., 2019a. vgraph: A generative model for joint community detection and node representation learning, in: Advances in Neural Information Processing Systems, pp. 514–524.

[45] Sun, Z., Deng, Z.H., Nie, J.Y., Tang, J., 2019b. Rotate: Knowledge graph embedding by relational rotation in complex space, in: International Conference on Learning Representations. URL: https://openreview.net/forum?id=HkgEQnRqYQ.

[46] Sun, Z., Vashishth, S., Sanyal, S., Talukdar, P., Yang, Y., 2019c. A re-evaluation of knowledge graph completion methods. arXiv preprint arXiv:1911.03903 .

[47] Tran, H.N., Takasu, A., 2019. Exploring scholarly data by semantic query on knowledge graph embedding space, in: International Conference on Theory and Practice of Digital Libraries, Springer. pp. 154–162.

[48] Trouillon, T., Welbl, J., Riedel, S., Gaussier, É., Bouchard, G., 2016. Complex embeddings for simple link prediction, in: International Conference on Machine Learning, pp. 2071–2080.

[49] Vu, T., Nguyen, T.D., Nguyen, D.Q., Phung, D., et al., 2019. A capsule network-based embedding model for knowledge graph completion and search personalization, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 2180–2189.

[50] Wang, H., Zhang, F., Zhang, M., Leskovec, J., Zhao, M., Li, W., Wang, Z., 2019a. Knowledge-aware graph neural networks with label smoothness regularization for recommender systems, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 968–977.

[51] Wang, K., Shen, Z., Huang, C., Wu, C.H., Dong, Y., Kanakia, A., 2020a. Microsoft academic graph: When experts are not enough. Quantitative Science Studies 1, 396–413.

[52] Wang, Q., Mao, Z., Wang, B., Guo, L., 2017. Knowledge graph embedding: A survey of approaches and applications. IEEE TKDE 29.

[53] Wang, W., Liu, J., Tang, T., Tuarob, S., Xia, F., Gong, Z., King, I., 2020b. Attributed collaboration network embedding for academic relationship mining. ACM Transactions on the Web (TWEB) 15, 1–20.

[54] Wang, Y., Yao, Q., Kwok, J.T., Ni, L.M., 2020c. Generalizing from a few examples: A survey on few-shot learning. ACM Computing Surveys (CSUR) 53, 1–34.

[55] Wang, Z., Ren, Z., He, C., Zhang, P., Hu, Y., 2019b. Robust embedding with multi-level structures for link prediction., in: IJCAI, pp. 5240–5246.

[56] Wolstencroft, K., Haines, R., Fellows, D., Williams, A., Withers, D., Owen, S., et al., 2013. The taverna workflow suite: designing and executing workflows of web services on the desktop, web or in the cloud. Nucleic acids research 41, W557–W561.

[57] Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., Philip, S.Y., 2020. A comprehensive survey on graph neural networks. IEEE Transactions on Neural Networks and Learning Systems .

[58] Yao, L., Zhang, Y., Wei, B., Jin, Z., Zhang, R., Zhang, Y., Chen, Q., 2017. Incorporating knowledge graph embeddings into topic modeling, in: Thirty-First AAAI Conference on Artificial Intelligence.

[59] Yin, W., 2020. Meta-learning for few-shot natural language processing: A survey. arXiv preprint arXiv:2007.09604 .

[60] Zhang, C., Yao, H., Huang, C., Jiang, M., Li, Z., Chawla, N.V., 2020. Few-shot knowledge graph completion, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 3041–3048.

[61] Zhang, S., Tay, Y., Yao, L., Liu, Q., 2019. Quaternion knowledge graph embedding. arXiv preprint arXiv:1904.10281 .

[62] Zhang, Y., Zhang, F., Yao, P., Tang, J., 2018. Name disambiguation in aminer: Clustering, maintenance, and human in the loop., in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1002–1011.

[63] Zhou, J., Cui, G., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., Sun, M., 2018. Graph neural networks: A review of methods and applications, in: CoRR.

[64] Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., He, Q., 2020. A comprehensive survey on transfer learning. Proceedings of the IEEE .