Link Prediction using Numerical Weights for Knowledge Graph Completion within the Scholarly Domain

Mojtaba Nayyeri[1], Gökce Müge Çil[1], Sahar Vahdati[2], Francesco Osborne[5], Andrey Kravchenko[3], Simone Angioni[6], Angelo Salatino[5], Diego Reforgiato Recupero[6], Enrico Motta[5], and Jens Lehmann[1,4]

[1] SDA Research Group, University of Bonn (Germany)
{nayyeri,cil,jens.lehmann}@cs.uni-bonn.de
[2] Nature-Inspired Machine Intelligence - InfAI, Dresden, Germany
vahdati@infai.org
[3] University of Oxford {andrey.kravchenko}@cs.ox.ac.uk
[4] Fraunhofer IAIS, Dresden (Germany) jens.lehmann@iais.fraunhofer.de
[5] Knowledge Media Institute, The Open University, Milton Keynes (UK)
{angelo.salatino, francesco.osborne, enrico.motta}@open.ac.uk
[6] Department of Mathematics and Computer Science, University of Cagliari (Italy)
{simone.angioni, diego.reforgiato}@unica.it

Abstract. Knowledge graphs (KGs) are widely used for modeling scholarly communication, informing scientometric analysis, and supporting a variety of intelligent services to explore the literature and predict research dynamics. However, they often suffer from incompleteness (e.g., missing affiliations, references, research topics), which ends up reducing the scope and quality of the resulting analysis. This issue is usually tackled by computing knowledge graph embeddings (KGEs) and applying link prediction techniques. However, only a few KGE models are capable of taking weights of facts in the knowledge graph into account. Such weights can have different meanings, e.g. describe the degree of association or the degree of truth. In this paper we propose *Weighted Triple Loss*, a new loss function for KGE models that takes full advantage of the additional numerical weights on facts. We also extend the *Rule Loss*, a loss function that is able to exploit a set of logical rules, in order to work with weighted triples. The evaluation of our solutions on several knowledge graphs indicates significant performance improvements with respect to the state of the art. Our main use case is the large-scale AIDA knowledge graph, which describes 21 million research articles. Our approach enables to complete information about affiliation types, countries, and research topics, greatly improving the scope of the resulting scientometrics analysis and providing better support to systems for monitoring and predicting research dynamics.

Keywords: Scholarly Data, Knowledge Graphs, Knowledge Graph Embeddings, Loss Functions, Link Prediction, Scholarly Communication, Science of Science.

1 Introduction

Science of Science is a rapidly emerging research field that studies the interactions among scientific agents in order to develop tools and policies for accelerating scientific process [9]. The large increase in the volume of scholarly outputs, such as articles, data sets, and software packages, yields unprecedented opportunities to this field, but also results in many challenges. This mass of available information has the potential to support a new generation of intelligent systems for exploring and improving research efforts, but at the same time poses a risk to drastically reduce the effectiveness of previous approaches for analysing available information. For instance, a recent article in Science [5] reported that the reaction to the COVID-19 pandemic is being slowed down by the fact that "scientists are drowning in COVID-19 papers" and need new solutions to efficiently analyse the scientific literature.

In order to address this challenge, we urge for structured, interlinked, and machine-readable representations of scholarly outputs. Knowledge Graphs (KGs) are becoming a standard solution for describing the actors (e.g., authors, organizations), the documents (e.g., publications, patents), and the research knowledge (e.g., research topics, tasks, technologies) in this space [13]. One of the main limitations of most KGs is the incompleteness problem, i.e., a large number of relevant facts are not present in the KG. Scholarly KGs are typically incomplete regarding crucial relations such as affiliations, references, research topics, conferences, and many others. This issue is usually tackled by producing a representation of the nodes and edges based on knowledge graph embeddings (KGEs) [15] and applying link prediction techniques [7] to this representation. Embedding models were successfully applied on KGs in different domains, including digital libraries [40], biomedical [18], and social media [33]. However, several KGs contain also facts with numerical weights in which the relationship is characterized by a numeric value, which is typically a confidence value, an intensity (e.g., the degree of association between an article and a research topic), or it further qualifies the information in the triple. Such a representation has already been described, analyzed and verified through a formal declarative semantics [37]. The resulting model theory is known as Annotated RDF (aRDF) and builds upon annotated logic. In aRDF any partially ordered set with a bottom element can be employed. For a given partially ordered set $(\mathcal{A}, \preceq)$, an element ϕ is the bottom iff $\phi \preceq x$ for all $x \in \mathcal{A}$. $\mathcal{A}$ might capture temporal, pedigree, possibilistic or fuzzy values. Since most of the existing KGE models can only handle triples that are either true or false, they perform quite poorly in this scenario as reported in Section 5.

In this paper, we propose the *Weighted Triple Loss*, a new loss function for KGE models that takes advantage of the additional numerical weights on facts. This loss can be used with different interaction models, e.g., DistMult [39], TransE [4], ComplEx [36]. We also extend the *Rule Loss* [22], a loss function that is able to exploit a set of automatically extracted logical rules, in order to work with weighted triples.

We implemented a KGE model based on DistMult which combines these two solutions and applied it on several knowledge graphs, obtaining significant performance improvements with respect to the state of the art.

The motivating scenario for this work concerns the Academia/Industry DynAmics (AIDA) Knowledge Graph [1], which was created for supporting an analysis of the flow of knowledge between academia and industry and systems for the prediction of research dynamics. The current version of AIDA integrates the metadata of about 21M research articles from Microsoft Academic Graph (MAG) and 8M patents from the Dimensions Dataset[7] in the field of Computer Science. AIDA classifies these documents according to the research topics from the Computer Science Ontology (CSO)[8] [29] and to the authors' affiliation types from the Global Research Identifier Database (GRID)[9] (e.g., 'education', 'company', 'government', 'healthcare'). This knowledge base enables tracking the evolution of research topics across academia, industry, government institutions, and other organizations. For instance, it was recently used for predicting the impact of specific research efforts on the industrial sector [28]. However, out of the 21M articles, only 5.1M were linked with GRID IDs in the source data and thus could be associated to their affiliation types and countries. Completing this data is thus crucial in order to improve the scope of different kinds of analysis about geopolitical factors [19], researcher migrations [20], collaboration patterns between academia and industry [2], and many others.

More in details, our main contributions are:

- Weighted Triple Loss, a loss function for weighted triples which is agnostic with respect to their meaning and takes advantage of the additional numerical weights on facts in order to handle KG incompleteness.
- An extension of the Rule Loss, which can handle weighted triples.
- *AIDA35k*[10], a new dataset describing 35K entities in the scholarly domain described by weighted triples.
- An evaluation showing that a KGE model based on DistMult that incorporates these loss functions outperforms several the state-of-the-art alternatives (UKGE, TransE, Distmult, and ComplEx) on AIDA35k, NL27K, CN15k and obtains competitive results on PPI5k.

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 the architecture of the new optimization technique. 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.

[7] Dimensions - `https://www.dimensions.ai/`

[8] CSO - `https://cso.kmi.open.ac.uk/`

[9] GRID - `https://www.grid.ac/`

[10] AIDA35k - `http://aida.kmi.open.ac.uk/aida35k/`

2 Related Work

A KGE model includes several components: embeddings (e.g., vector, matrix, tensor), a score function, and a loss function. The score function $f(h,r,t)$ takes as input the embeddings of a triple $(\mathbf{h},\mathbf{r},\mathbf{t})$ and returns a value reflecting how plausible the triple is. These scores depend on the optimization of the loss function. Most modern KGE models either use distance functions (e.g., TransE [4], RotatE [34]) or inner product functions based on semantic matching (e.g., DistMult [39], ComplEx [36], QuatE [41], RESCAL [24]).

Typically KGE models employ loss functions that can be applied only on positive and negative triples, such as as Margin Ranking Loss (MRL) [4], negative likelihood of logistic model [36], self adversarial loss [34], and full multiclass log loss [17]. An exception is RESCAL [24] which adopts the Mean Square Error (MSE) loss and thus can be trained also on weighted triples, where the weight reflects the uncertainty of a triple. A more recent work [6] combines the MSE loss with a rule-based loss in order to improve the ability to learn from weighted triples. In the next section we will review these loss functions in detail.

2.1 Loss Functions for KGEs

Margin Ranking Loss [4]. The margin ranking loss which is inspired by general margin based approaches [3] aims at forcing a margin between each positive sample (h,r,t) and its corresponding negative samples (h',r,t'). The negative samples are generated by replacing either head or tail of positive samples with a random entity from the KG. The formulation of the MRL is $\mathcal{L}_{MRL} = \sum_{(h,r,t)\in\mathcal{T}}\sum_{(h',r,t')\in\mathcal{N}_{h,r,t}}[f(h,r,t)+\gamma-f(h',r,t')]_+$, where $[x]_+ = \max(0,x)$, $\mathcal{T}$ is the set of all positive samples, $\mathcal{N}^-_{h,r,t}$ is the set of all negative samples generated from the triple (h,r,t), and γ is the length of the margin between positive and negative samples.

Limit-based Scoring Loss [42]. When a KGE model is trained by using the MRL, the score of positive triples may be unbounded. In the case of translation based KGE model, such a limitation would prevent the model from fulfilling the translation in the vector space, resulting in poor performance [42]. The limit-based scoring loss aims at avoiding this issue by including the boundary for the scores of positive triples in the margin ranking loss.

$$\mathcal{L}_{LSL} = \sum_{(h,r,t)\in\mathcal{T}}\sum_{(h',r,t')\in\mathcal{N}_{h,r,t}}[f(h,r,t)+\gamma-f(h',r,t')]_+ + \lambda[f(h,r,t)-\gamma_1]_+ \tag{1}$$

The term $[f(h,r,t)-\gamma_1]_+$ enforces the constraint $f(h,r,t)\leq\gamma_1$. Therefore, scores of positive triples are bounded not to be very large.

Soft Margin Loss (SML) [21]. The soft margin loss aims to handling noisy negative samples. It adds slack variables to negative samples optimization in order to mitigate the negative effect of false negative samples

$$\mathcal{L}_{SML} = \sum_{(h,r,t)\in\mathcal{T}} \sum_{(h',r,t')\in\mathcal{N}_{h,r,t}} \lambda\, \eta^2_{h,r,t} + \lambda_+[f(h,r,t) - \gamma_1]_+ + \lambda_-\,[\gamma_2 - f(h',r,t') - \eta_{h,r,t}]_+ \tag{2}$$

SlidE Loss [23]. The length of the margin is an important factor which affects the performance of KGE models. In the above mentioned loss functions, the length of the margin is obtained by trial and error. In Limit-based score loss and Soft Margin Loss, the length of the margin is determined by two hyper-parameters. Therefore, the search space is large. The SlidE loss [23] addresses this problem by determining the center of the margin and automatically adjusting the length of the margin by means of a trainable variable. The formulation of the SlidE loss is as follows

$$\mathcal{L}_{SlidE^+} = \lambda e^{-\sigma\eta^2} + \lambda_+\,[f(h,r,t) - \gamma + \eta]_+ + \lambda_-\,[-f(h^{'},r,t^{'}) + \gamma + \eta]_+ \tag{3}$$

Self Adversarial Loss (SAL) [34]. Self Adversarial Loss obtained state of the art performance on distance based KGE models. The formulation of the loss is

$$\mathcal{L} = -\sum_{(h,r,t)\in\mathcal{T}} \Big(\log\sigma(\gamma - f(h,r,t)) + \sum_{(h',r,t')\in\mathcal{N}_{h,r,t}} p(h',r,t')\log\sigma(f(h',r,t') - \gamma)\Big), \tag{4}$$

where $p(h',r,t') = \frac{exp(\alpha f(h',r,t'))}{\sum exp(\alpha f(h',r,t'))}$, α is adversarial temperature. The loss puts the higher weight for negative samples to reduce the score of negative samples with higher scores during the training process.

Negative Log Likelihood Loss (NLL) [36]. The negative likelihood of the logistic model with regularization is as follows

$$\mathcal{L}_{NLL} = \sum_{(h,r,t)\in\mathcal{T}\cup\mathcal{N}} \lambda\, \log(1 + exp(-y_{h,r,t} f(h,r,t))) + \lambda||\theta||^2, \tag{5}$$

where θ is all adjustable parameters.

Full Multiclass Log Loss (FMLL) [17]. The ComplEx model was originally trained using the negative likelihood log loss. However, it has been recently shown that the model obtains state-of-the-art performance by using the full multiclass log-loss. The loss applies full negative sampling and is defined as follows

$$\mathcal{L}_{FMLL} = \sum_{(h,r,t)\in\mathcal{T}} l(f(h,r,t))), \tag{6}$$

where $l(f(h,r,t))) = l^1(f(h,r,t)) + l^2(f(h,r,t)),\ \ l^1(f(h,r,t)) = -f(h,r,t) + log(\sum_{t'} exp(f(h,r,t'))),\ \ l^2(f(h,r,t)) = -f(h,r,t) + log(\sum_{h'} exp(f(h',r,t)))$.

The loss gives big (small) scores for positive (negative) triples.

2.2 Loss For Weighted Triples

The mentioned loss functions are suitable for learning over triples which are either positive or negative. Recently, the MSE loss, already used for training the RESCAL model, has been used for training over KGs with uncertain triples i.e., triples associated a weight that reflects the confidence in their correctness [6]. This loss is model independent and it is formulated as in the following:

$$\mathcal{L}_{UKGE} = \sum_{(h,r,t)\in\mathcal{T}_w\cup\mathcal{N}} |f(h,r,t) - w_{h,r,t}|^2 + \sum_{(h,r,t)} \sum_{g} |\psi_g(f(h,r,t))|^2, \qquad (7)$$

where $w_{h,r,t}$ is the weight of a triple (h, r, t) and $\mathcal{T}_w$ is the set of all weighted triples. g refers to a rule and ψ_g is the weighted distance of the rule g obtained by probabilistic soft logic (psl). We hypothesize that the weights might not be exact (due to noise, KG incompleteness, and so on). Therefore, relaxing the model by adding complementary variables that enable to learn the weight by a tolerance will be consistent with the nature of weighted triples and may improve the performance.

2.3 Scholarly Knowledge Graphs

Knowledge graphs about research outputs typically either focus on the metadata (e.g., titles, abstracts, authors, organizations) or they offer a machine-readable representation of the knowledge contained in research articles.

A good example of the first category is Microsoft Academic Graph (MAG) [38], 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[11] 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 [27] 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 Open Academic Graph (OAG)[12] is a large knowledge graph integrating 208M papers from MAG and 172M from AMiner.

All these resources suffer from data incompleteness to different degrees. For instance, it is still challenging to identify and disambiguate affiliations, which hinders our ability to categorize the articles according to their affiliation type or country [19]. Similarly, references are usually incomplete, and the citation count of the same paper tends to vary dramatically on different datasets [27].

A second category of knowledge graphs focuses instead on representing the content of scientific publications. This objective was traditionally pursued by the semantic web community, e.g., by creating bibliographic repositories in the Linked Data Cloud [25], encouraging the Semantic Publishing paradigm [32],

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

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

implementing systems for managing nano-publications [11,16] and micropublications [31], and developing a variety of ontologies to describe scholarly data, e.g., SWRC[13], BIBO[14], BiDO[15], SPAR [26][16], CSO [30]. A recent project is the Open Research Knowledge Graph (ORKG) [14], 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)[17] 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.

3 Motivating Scenario: the AIDA Knowledge Graph

New scientific knowledge is continuously produced by the collective effort of a variety of actors, such as universities, commercial companies, government institutions, non-profit, and many others. Analysing how these organizations collaborate in different research areas and exchange ideas and persons is crucial for harmonising their efforts as well as for understanding, monitoring, and anticipating research dynamics [2].

In order to support such analysis, we recently released the Academia/Industry DynAmics (AIDA) Knowledge Graph [1], a resource that includes more than one billion triples and describes 21M publications from Microsoft Academic Graph (MAG)[18] [38] and 8M patents from Dimensions. 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/`). All the articles and patents in AIDA are associated with a distribution of topics from the Computer Science Ontology (CSO) [29], which is the largest taxonomy in the field, counting more than 14K topics. 5.1M publications and 5.6M patents are also categorized according to the type of the author's affiliations from the Global Research Identifier Database (GRID), a openly accessible database of research institution identifiers. The classification is composed by eight exclusive categories: Education, Healthcare, Company, Archive, Nonprofit, Government, Facility, and Other.

The combination of organization types and topics in AIDA allows researchers to produce several kinds of advanced analysis. For instance, it was recently used to improve the state of the art regarding the prediction of research impact on the industrial sectors [28].

Table 1 shows, as example, the number of publications in three well known research topics classified according to the percentage of authors in organization type (we show just five for space constraints). For instance, about 15.7K of the

[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] AI-KG - `http://scholkg.kmi.open.ac.uk/`

[18] MAG - `https://academic.microsoft.com/`

Neural Networks articles have at least one author from a company, 11.7K have at least half of the authors in this category, and only 8.6K have all the authors. Overall this data show that these organization types are very different in term of contributions. Authors from academia tend mostly to collaborate among themselves, and the same is true even if to a lesser degree for authors from companies. Conversely, the other categories tend to collaborate more with different types. Only 8.1% (3,969,096-3,648,629) of the Computer Science papers involving at least an author from Education includes also one or more different categories (e.g., Company, Government). This number raises to 14.6% for Company, 46.1% for Government, 46.6% for Nonprofit, and 69.0% for Healthcare. However, these dynamics can vary drastically in different research areas. For instance, companies tend to collaborate much more with other categories (mostly universities under Education) in the fields of Neural Networks (45.7% of collaborations) and Semantic Web (49.8%).

Table 1: AIDA - Number of papers in a topic with an organization type according to the percentage of authors in the category (> 0, ≥ 0.5, $= 1.0$)

Topic	**Education**	**Company**	**Government**	**Nonprofit**	**Healthcare**
Computer Science (> 0)	3,969,097	954,143	185,633	61,129	15,163
Computer Science (≥ 0.5)	3,895,432	877,021	140,889	45,049	8,265
Computer Science ($= 1.0$)	3,648,629	814,610	100,100	32,619	4,696
Neural Networks (> 0)	219,492	15,776	9,918	2,336	1,660
Neural Networks (≥ 0.5)	215,146	11,761	7,163	1,554	727
Neural Networks ($= 1.0$)	202,161	8,565	4,755	1,065	347
Semantic Web (> 0)	38,306	2,703	2,205	1,018	285
Semantic Web (≥ 0.5)	37,554	1,888	1,628	720	178
Semantic Web ($= 1.0$)	34,780	1,358	1,094	493	95

The main shortcoming of the current version of AIDA is that only about 25% of the articles (5.1M out of 21M) and 70% of the patents (5.6M out of the 8M) are associates to the GRID affiliation type. The missing data are due to the fact that some affiliations are not present on GRID or they were not correctly mapped to the relevant GRID IDs in the original data. In order to improve the scope of the analyses supported by AIDA is thus critical to address this issue by mapping articles to the correct organization type.

This scenario motivated us to investigate different models for link prediction that could be applied on AIDA and on other knowledge graphs that suffer from the same issues. However, as previously mentioned, several information regarding the documents in AIDA are best represented as weighted triples. For instance, since the authors of a paper can have different affiliation types, each category is associated to a weight equal to the fraction of authors associated with that type. Therefore, a paper that has three authors associated with the type ‘Education’ and one with the type ‘Industry’ would be assigned the category ‘Education’ with a weight of 0.75 and the category ‘Company’ with a weight of 0.25. This can be represented as two weighted triples: $< paperID, hasGridType, Education, 0.75 >$ and $< paperID, hasGridType, Company, 0.25 >$. The same mechanism is also used to associate articles to countries: a paper that has half of the authors from UK will be associated with the weighted triple: $< paperID, hasCountry, UK, 0.5 >$. This same solution is also used to quantify the number of citations received by a paper in a specific year. When representing these data as Resource Description Framework (RDF) we need to reify these triples as shown by Figure 1.

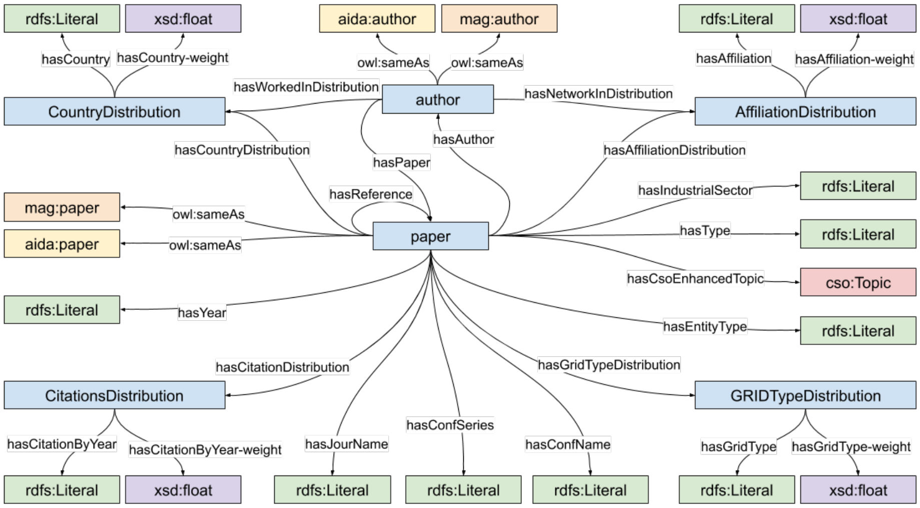

Fig. 1: RDF Schema for articles in AIDA.

These considerations led to the design of the loss functions presented in this paper. In order to complete AIDA KG, we implemented a version of DistMult that incorporates the Weighted Triple loss function, labelled in the following *Weighted Graph Embedding* (WGE).

In order to empirically evaluate the effectiveness of our loss function, we apply it on the AIDA knowledge graph by generating a subset named AIDA35k, a new dataset including 35K entities from AIDA associated to triples with numerical weights. AIDA35k is a weighted Knowledge Graph $\mathcal{K}$, where $\mathcal{K} = \{\mathcal{E}, \mathcal{R}, \mathcal{T}_w\}$ and $\mathcal{T}_w = \{(h, r, t, w_{h,r,t})\} \subset \mathcal{E} \times \mathcal{R} \times \mathcal{E} \times \mathbb{R}$. $w_{h,r,t}$ is the weight of the fact (h, r, t).

4 Optimising KGE models for Weighted Triples

In this section, we propose two loss functions: Weighted Triple Loss and Rule Loss for Weighted Triples. Those loss functions optimise weighted triples of the form $(h, r, t, w_{h,r,t})$ where h and t are head and tail entities, r is a relation between them, and $w_{h,r,t}$ is the weight assigned to the triple (h, r, t).

4.1 Weighted Triple Loss

The loss function is agnostic with respect to the kind of weight. Conceptually, we consider two main types of weights. The first is related to the *correctness* of the triples and indicates its degree of plausibility. The second refers to the *intensity* of the relation and reflects the degree of association between the head and the tail.

The main intuition behind the Weighted Triple Loss is that in many practical cases $w_{h,r,t}$ is estimated with a certain degree of uncertainty. We model this uncertainty as follows:

$$w_{h,r,t} - \eta_{h,r,t}^{-^2} \leq f(h,r,t) \leq w_{h,r,t} + \eta_{h,r,t}^{+^2}, \tag{8}$$

where $f(h,r,t)$ is the score of the triple (h,r,t) given by a KGE interaction model. $\eta_{h,r,t}^{-^2}$ and $\eta_{h,r,t}^{+^2}$ are trainable variables which allow the score $f(h,r,t)$ not to be exactly equal to $w_{h,r,t}$, but rather to be a number bounded between $w_{h,r,t} - \eta_{h,r,t}^{-^2}$ and $w_{h,r,t} + \eta_{h,r,t}^{+^2}$. In order to optimize embedding vectors of entities and relations as well as adjusting $\eta_{h,r,t}^{-^2}, \eta_{h,r,t}^{+^2}$, the following optimization framework is proposed:

$$\begin{cases} \min_\theta \sum_{(h,r,t,w_{h,r,t}) \in \{\mathcal{T}_w \cup \mathcal{N}\}} \lambda_1 \eta_{h,r,t}^{-^2} + \lambda_2 \eta_{h,r,t}^{+^2} + \lambda_3 \mathcal{L}, \text{ s.t.} \\ w_{h,r,t} - \eta_{h,r,t}^{-^2} \leq f(h,r,t) \leq w_{h,r,t} + \eta_{h,r,t}^{+^2}, \end{cases} \tag{9}$$

where λ_1, λ_2 are hyper-parameters that affect the degree to which $\eta_{h,r,t}^{-}, \eta_{h,r,t}^{+}$ are minimized, λ_3 is the multiplier of regularization term over embedding of entities ad relations, θ contains all adjustable parameters including embeddings of entities and relations and $\eta_{h,r,t}^{-^2}, \eta_{h,r,t}^{+^2}$ i.e. $\theta = \{(\mathbf{h}, \mathbf{r}, \mathbf{t}) \cup \{\eta_{h,r,t}^{-^2}, \eta_{h,r,t}^{+^2}\} | (h,r,t) \in \mathcal{T}\}$. $\mathcal{L}$ is the regularization over entities and relations embedding i.e. $\mathcal{L} = \mathbf{E^2} + \mathbf{R^2}$. $\mathbf{E}$ and $\mathbf{R}$ are the embeddings of all entities and relations in the KG. For each quadruple in the training set $(h,r,t,w_{h,r,t})$, we generate a corrupted sample using uniform negative sampling technique [4] where either h or t is replaced by a random entity $e \in \mathcal{E}$, i.e., the resulting triples are $(h' = e, r, t, w_{h',r,t})$ or $(h, r, t', w_{h,r,t'})$. For corrupted samples, we set their weights w to zero. We indicate the set of all corrupted samples by $\mathcal{N}$.

4.2 Rule Loss for Weighted Triples

Extraction of Rules. In order to include additional logical rules as complementary knowledge, we used the AMIE rule extractor [10] which is specifically designed for rule extraction on KGs. A logical rule is generally of the form of $PREMISE \xrightarrow{PCA} CONCLUSION$ where $PREMISE$ can be constructed from different relations with joint head or tail. For instance, the probability of the rule $?e$ *hasAuthor* $?a$ AND $?e$ *hasCountry* $?b \xrightarrow{0.553} ?a$ *workedIn* $?b$ is 55.3% which is assigned by AIME.

Definition of the Rule Loss for Weighted Triples. In order to apply rules to weighted triples we extend the approach presented in Nayyeri et al. [22]. For a given rule of the form $rule : q_1 \wedge q_2 \wedge \ldots \wedge q_n \rightarrow q_{n+1}$ where q_i, $\imath = 1, \ldots, n+1$ are atoms (weighted triples where relations are fixed, but entities are variable). To model rule loss for the above-mentioned rule, we use the following formula.

$$\mathcal{R} = \max(w_{q_1} * \ldots * w_{q_n} - f(q_{n+1}), 0), \tag{10}$$

where w_{q_i} is the weight of the weighted triples $q_i, i = 1, \ldots, n$ after grounding of entities (replacing variables by entities in $\mathcal{E}$). $f(q_{n+1})$ is the score of the triple (h, r, t) in the weighted triples $(h, r, t, w_{h,r,t})$ where $w_{h,r,t}$ is not given in the training set, but is approximated by the score of the used KGE model i.e., $f(q_{n+1}) = f(h, r, t)$. For each $rule_i, i = 1, \ldots, l$ in the rule set, we provide the corresponding rule loss $\mathcal{R}_i$. The rule loss can be added to the optimization framework as

$$\begin{cases} \min_\theta \sum_{(h,r,t,w_{h,r,t}) \in \{\mathcal{T}_w \cup \mathcal{N}\}} \; \lambda_1 \eta^{-^2}_{h,r,t} + \lambda_2 \eta^{+^2}_{h,r,t} + \lambda_3 \mathcal{L} + \lambda_4 \sum_{i=1}^{l} \mathcal{R}_i, \text{ s.t.} \\ w_{h,r,t} - \eta^{-^2}_{h,r,t} \leq f(h,r,t) \leq w_{h,r,t} + \eta^{+^2}_{h,r,t}, \end{cases} \tag{11}$$

or added as additional weighted triples $\mathcal{T}'_w = \{(h, r, t, w_{h,r,t} = w_{q_1} * \ldots * w_{q_n})\}$, where (h, r, t) is in the head of a rule $q_1 \wedge q_2 \wedge \ldots \wedge q_n \rightarrow (h, r, t)$. Therefore, the following optimization problem is suggested

$$\begin{cases} \min_\theta \sum_{(h,r,t,w_{h,r,t}) \in \{\mathcal{T} \cup \mathcal{T}'_w \cup \mathcal{N}\}} \; \lambda_1 \eta^{-^2}_{h,r,t} + \lambda_2 \eta^{+^2}_{h,r,t} + \lambda_3 \mathcal{L}, \text{ s.t.} \\ w_{h,r,t} - \eta^{-^2}_{h,r,t} \leq f(h,r,t) \leq w_{h,r,t} + \eta^{+^2}_{h,r,t}. \end{cases} \tag{12}$$

5 Evaluation

In this section, we compare the performance of i) Weighted Graph Embedding (*WGE*), the version of DistMult that incorporates the Weighted Triple loss function (see Section 3), ii) the Uncertain KG Embedding (*UKGE*), which uses the loss function presented in Chen at al. [6] (see Section 2.2), iii) *DistMult*[39], iv) *TransE*[4], and v) *ComplEx*[36] on several datasets. The evaluation data are available at `http://aida.kmi.open.ac.uk/aida35k/`.

We consider two different versions of the WGE and UKGE models namely rectified (tagged as *rect*), and logistic (*logi*) [6]. These functions maps the plausibility of a triple according to a score between 0 and 1: the logistic function uses a smooth approach while the rectifier hardly clamps the values. On AIDA, we further distinguish between the two versions of WGE that use the Rules Loss defined in Section 4.2 and the ones that do not.

5.1 Experimental Setup

Dataset. In addition to AIDA35k, which was introduced in Section 3, we used three other datasets that include weighted triples: CN15k, NL27k, and PPI5k. These datasets were used in the evaluation of UKGE [6], which is one of the baselines. They were originally extracted from ConceptNet, NELL, and the Protein-Protein Interaction Knowledge Base STRING [35].

Rule Extraction We set a probability threshold of 0.4 for the extracted AMIE rules. When binding rule variables to entities, those rules generate *grounded*

Table 2: Statistical information about the datasets. Avg(s) and Std(s) are the average and standard deviation of the scores.

Dataset	No. Entities	No. Relations	No.Triples	Avg(s)	Std(s)
AIDA35k	35,229	17	183,601	0.906	0.261
CN15k	15,000	36	241,158	0.629	0.232
PPI5k	4,999	7	271,666	0.415	0.213
NL27k	27,221	404	175,412	0.797	0.242

triples. We used a threshold of 0.1 for filtering those grounding triples. Overall, this process generated 18 rules and 126,031 grounded triples. Among these, 20,450 grounding triples belong to *hasGridType* relation.

Table 3: Selected examples of extracted rules in AIDA and statistics.

Example Rule	Prob.	Triples
?a hasAuthor ?b → *?b hasPaper ?a*	1	1,034
?f hasEntityType ?a AND ?b hasReference ?f →*?b hasEntityType ?a*	1	770
?f hasGridType ?a AND ?b hasReference ?f → *?b hasGridType ?a*	0.6	1,273
?f hasCountry ?a AND ?b hasPaper ?f → *?b workedIn ?a*	0.5	4,055

Metrics and Hyperparameters. We adopted the Mean Square Error (MSE), the Mean Absolute Error (MAE), and the F1 measure as evaluation metrics. Since the space is limited and they are standard metrics used by most works in this field [12], they will not be described in this paper.

The set up of the experiments includes the sets of hyperparameters with batch sizes $\{256, 512, 1024\}$, and learning rate of $\{0.1, 0.01, 0.001, 0.0001\}$. The embedding dimension is $\{64, 128, 256, 512\}$ with 10 negative sampling. The regularization scale for the rectified versions is $\{0.5, 0.1, 0.05, 0.01, 0.005, 0.001\}$. The rule coefficient in rule loss is trained for $\{0.1, 0.5, 1.0\}$.

5.2 Results

Table 4 reports the performance of the approaches on the four datasets. On the AIDA35k dataset, the logistic version of our model with rules outperforms all alternatives according to the MSE and MAE metrics.

The rectified version of our model obtains the best results in terms of F1 (0.847) and Accuracy (0.871). On the NL27K dataset, the rectifier version without rules (*WGE_rect_no_rules*) outperforms the other approaches in MSE and MAE, while the logistic versions of our model (*WGE_rect_no_rules*) achieves the best results in F1 and accuracy. On PPI5k, WGE obtains competitive results, outperforming TransE, Distmult, and Complex in all the metrics and UKGE in MAE. However, UKGE performs better in F1 and MSE and yields comparable accuracy. This is due to the limited size of PPI5k which includes only 5K entities and 7 relations. Finally, on CN15k the rectified WGE without rules (*WGE_rect_no_rules*) reaches the highest performance in all the metrics.

Generally, the models which include the weighted triple loss outperform those with normal loss by a large margin. In particular, the difference in terms of F1 score and accuracy between the standard DistMult model and the DistMult

Table 4: Performance of approaches on the four benchmarks (AIDA35k, NL27K, PPI5k, CN15k). In bold the best results.

	MSE	**MAE**	**F1**	**Accuracy**	**MSE**	**MAE**	**F1**	**Accuracy**
	AIDA35K				***NL27K***			
UKGE_rect	0.215	0.299	0.698	0.795	0.027	0.073	0.929	0.956
UKGE_logi	0.234	0.341	0.699	0.731	0.055	0.111	0.904	0.942
WGE_rect_no_rules	0.143	0.179	0.842	0.865	**0.019**	**0.065**	0.901	0.934
WGE_logi_no_rules	0.874	0.129	0.750	0.743	0.032	0.076	**0.937**	**0.960**
WGE_rect_with_rules	0.135	0.170	**0.847**	**0.871**	-	-	-	-
WGE_logi_with_rules	**0.097**	**0.128**	0.808	0.820	-	-	-	-
TransE	-	-	0.674	0.642	-	-	0.651	0.534
DistMult	-	-	0.741	0.642	-	-	0.721	0.701
ComplEx	-	-	0.748	0.778	-	-	0.633	0.534
	PPI5K				***CN15K***			
UKGE_rect	**0.003**	0.028	**0.969**	**0.997**	0.161	0.322	0.259	0.902
UKGE_logi	0.005	0.035	0.967	**0.997**	0.281	0.448	0.258	0.901
WGE_rect_no_rules	0.004	**0.022**	0.912	0.992	**0.148**	**0.307**	**0.304**	**0.902**
WGE_logi_no_rules	0.009	0.049	0.921	0.993	0.223	0.388	0.279	0.861
TransE	-	-	0.832	0.985	-	-	0.234	0.679
DistMult	-	-	0.869	0.979	-	-	0.279	0.711
Complex	-	-	0.832	0.989	-	-	0.189	0.732

interaction model with the best variant of our proposed loss function is higher than 10% on average across all datasets. This empirically confirms our research hypothesis that there is a substantial benefit in using triple weight information – if available in the KG – in the loss function.

In Table 5, we present a relation-specific evaluation considering *hasGridType* and *hasCountry* from AIDA35k. For *hasGridType*, *WGE_rect_no_rules* achieves the best results for MAE, F1, and accuracy. For *hasCountry*, the rectified version of WGE with rules outperforms all the other models in MSE, F1, and accuracy. This confirms the usefulness of rules for handling specific relations.

6 Conclusions and Future Work

In this paper we proposed Weighted Triple Loss, a new loss function for KGE models that takes full advantage of the additional numerical weights on facts in order to handle KG incompleteness. We also extend the Rule Loss, a loss function that is able to exploit a set of logical rules, in order to work with weighted triples. The resulting KGE model, labelled Weighted Graph Embedding (WGE), was designed to address the real word scenario of completing the AIDA Knowledge Graph, in order to enable more comprehensive quantitative analysis of science about geopolitical factors [19] and the flow of knowledge between different types of organizations [2] (e.g., university, industry, non-profit). However, the loss functions are general solutions that can be combined with any interaction model in order to take into account weighted triples. The evaluation

Table 5: Performance when considering only the relations *hasGridType* and *hasCountry* in AIDA35k. In bold the best results.

	MSE	**MAE**	**F1**	**Accuracy**	**MSE**	**MAE**	**F1**	**Accuracy**
	AIDA35k (hasGridType)				***AIDA35k (hasCountry)***			
UKGE_rect	0.057	0.174	0.714	0.824	0.104	0.268	0.318	0.797
UKGE_logi	0.122	0.298	0.793	0.835	0.158	0.328	0.574	0.723
WGE_rect_no_rules	0.049	**0.149**	**0.848**	**0.882**	0.057	**0.177**	0.696	0.790
WGE_logi_no_rules	**0.045**	0.152	0.694	0.702	0.136	0.243	0.613	0.689
WGE_rect_with_rules	0.060	0.175	0.833	0.864	**0.053**	0.183	**0.713**	**0.804**
WGE_logi_with_rules	0.046	0.177	0.649	0.651	0.130	0.242	0.629	0.709
DistMult	-	-	0.562	0.518	-	-	0.621	0.736
TransE	-	-	0.474	0.445	-	-	0.530	0.574
ComplEx	-	-	0.786	0.824	-	-	0.575	0.716

showed that our solutions outperform the baselines (UKGE, TransE, Distmult, and ComplEx) on AIDA35k, NL27K, CN15k and obtain competitive results on PPI5k. The approach presented in this paper opens up several interesting directions of work. First, we aim to apply our approach on other KGs in this space for improving their coverage of the research landscape. We also plan to run it on recent KGs describing scientific concepts (e.g., tasks, methods, materials) and their relationships, such as AI-KG [8] and ORKG [14], where the numerical weights could represent the consensus of the research community on the relevant statements. Finally, we intend to study the application of our solutions on approaches for classifying research articles and predicting research trends.

References

1. Angioni, S., Salatino, A.A., Osborne, F., Recupero, D.R., Motta, E.: Integrating knowledge graphs for analysing academia and industry dynamics. In: ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium. pp. 219–225. Springer (2020)
2. Ankrah, S., Omar, A.T.: Universities–industry collaboration: A systematic review. Scandinavian Journal of Management **31**(3), 387–408 (2015)
3. Awad, M., Khanna, R.: Support vector regression. In: Efficient learning machines, pp. 67–80. Springer (2015)
4. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., Yakhnenko, O.: Translating embeddings for modeling multi-relational data. In: Advances in NIPS (2013)
5. Brainard, J.: Scientists are drowning in covid-19 papers. can new tools keep them afloat. Science (2020)
6. Chen, X., Chen, M., Shi, W., Sun, Y., Zaniolo, C.: Embedding uncertain knowledge graphs. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 3363–3370 (2019)
7. Dai, Y., Wang, S., Xiong, N.N., Guo, W.: A survey on knowledge graph embedding: Approaches, applications and benchmarks. Electronics **9**(5), 750 (2020)
8. Dessı, D., Osborne, F., Recupero, D.R., Buscaldi, D., Motta, E., Sack, H.: Ai-kg: an automatically generated knowledge graph of artificial intelligence

9. Fortunato, S., Bergstrom, C.T., Börner, K., Evans, J.A., Helbing, D., Milojević, S., Petersen, A.M., Radicchi, F., Sinatra, R., Uzzi, B., et al.: Science of science. Science **359**(6379), eaao0185 (2018)
10. Galárraga, L., Teflioudi, C., Hose, K., Suchanek, F.M.: Fast rule mining in ontological knowledge bases with AMIE+. The VLDB Journal – The International Journal on Very Large Data Bases **24**(6), 707–730 (2015)
11. Groth, P., Gibson, A., Velterop, J.: The anatomy of a nanopublication. Information Services & Use **30**(1-2), 51–56 (2010)
12. Gunawardana, A., Shani, G.: A survey of accuracy evaluation metrics of recommendation tasks. Journal of Machine Learning Research **10**(12) (2009)
13. Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., de Melo, G., Gutierrez, C., Gayo, J.E.L., Kirrane, S., Neumaier, S., Polleres, A., et al.: Knowledge graphs. arXiv preprint arXiv:2003.02320 (2020)
14. Jaradeh, M.Y., Oelen, A., Farfar, K.E., Prinz, M., D'Souza, J., Kismihók, G., Stocker, M., Auer, S.: Open research knowledge graph: Next generation infrastructure for semantic scholarly knowledge. In: Proceedings of the 10th International Conference on Knowledge Capture. pp. 243–246 (2019)
15. Ji, S., Pan, S., Cambria, E., Marttinen, P., Yu, P.S.: A survey on knowledge graphs: Representation, acquisition and applications. arXiv preprint arXiv:2002.00388 (2020)
16. Kuhn, T., Chichester, C., Krauthammer, M., Queralt-Rosinach, N., Verborgh, R., Giannakopoulos, G., Ngomo, A.C.N., Viglianti, R., Dumontier, M.: Decentralized provenance-aware publishing with nanopublications. PeerJ Computer Science **2**, e78 (2016)
17. Lacroix, T., Usunier, N., Obozinski, G.: Canonical tensor decomposition for knowledge base completion. In: International Conference on Machine Learning. pp. 2863–2872 (2018)
18. Li, L., Wang, P., Wang, Y., Jiang, J., Tang, B., Yan, J., Wang, S., Liu, Y.: Prtransh: Embedding probabilistic medical knowledge from real world emr data. arXiv preprint arXiv:1909.00672 (2019)
19. Mannocci, A., Osborne, F., Motta, E.: Geographical trends in academic conferences: An analysis of authors' affiliations. Data Science **2**(1-2), 181–203 (2019)
20. Moed, H.F., Aisati, M., Plume, A.: Studying scientific migration in scopus. Scientometrics **94**(3), 929–942 (2013)
21. Nayyeri, M., Vahdati, S., Zhou, X., Yazdi, H.S., Lehmann, J.: Embedding-based recommendations on scholarly knowledge graphs. In: European Semantic Web Conference. pp. 255–270. Springer (2020)
22. Nayyeri, M., Xu, C., Lehmann, J., Yazdi, H.S.: Logicenn: A neural based knowledge graphs embedding model with logical rules. arXiv preprint arXiv:1908.07141 (2019)
23. Nayyeri, M., Zhou, X., Vahdati, S., Izanloo, R., Yazdi, H.S., Lehmann, J.: Let the margin slide±for knowledge graph embeddings via a correntropy objective function. In: 2020 International Joint Conference on Neural Networks (IJCNN). pp. 1–9. IEEE (2020)
24. Nickel, M., Tresp, V., Kriegel, H.P.: A three-way model for collective learning on multi-relational data. In: ICML. vol. 11, pp. 809–816 (2011)
25. Nuzzolese, A.G., Gentile, A.L., Presutti, V., Gangemi, A.: Semantic web conference ontology-a refactoring solution. In: European Semantic Web Conference. pp. 84–87. Springer (2016)
26. Peroni, S., Shotton, D.: The spar ontologies. In: International Semantic Web Conference. pp. 119–136. Springer (2018)

27. Peroni, S., Shotton, D.: Opencitations, an infrastructure organization for open scholarship. Quantitative Science Studies **1**(1), 428–444 (2020)
28. Salatino, A., Osborne, F., Motta, E.: Researchflow: Understanding the knowledge flow between academia and industry. In: Knowledge Engineering and Knowledge Management – 22nd International Conference, EKAW 2020 (2020)
29. Salatino, A.A., Thanapalasingam, T., Mannocci, A., Birukou, A., Osborne, F., Motta, E.: The computer science ontology: A comprehensive automatically-generated taxonomy of research areas. Data Intelligence **2**(3), 379–416 (2020)
30. Salatino, A.A., Thanapalasingam, T., Mannocci, A., Osborne, F., Motta, E.: The computer science ontology: a large-scale taxonomy of research areas. In: ISWC. pp. 187–205 (2018)
31. Schneider, J., Ciccarese, P., Clark, T., Boyce, R.D.: Using the micropublications ontology and the open annotation data model to represent evidence within a drug-drug interaction knowledge base (2014)
32. Shotton, D.: Semantic publishing: the coming revolution in scientific journal publishing. Learned Publishing **22**(2), 85–94 (2009)
33. Stanovsky, G., Gruhl, D., Mendes, P.: 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 (2017)
34. Sun, Z., Deng, Z.H., Nie, J.Y., Tang, J.: Rotate: Knowledge graph embedding by relational rotation in complex space. In: International Conference on Learning Representations (2019)
35. Szklarczyk, D., Morris, J.H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N.T., Roth, A., Bork, P., et al.: The string database in 2017: quality-controlled protein–protein association networks, made broadly accessible. Nucleic acids research p. gkw937 (2016)
36. Trouillon, T., Welbl, J., Riedel, S., Gaussier, É., Bouchard, G.: Complex embeddings for simple link prediction. In: ICML (2016)
37. Udrea, O., Recupero, D.R., Subrahmanian, V.S.: Annotated RDF. ACM Trans. Comput. Log. **11**(2), 10:1–10:41 (2010)
38. Wang, K., Shen, Z., Huang, C., Wu, C.H., Dong, Y., Kanakia, A.: Microsoft academic graph: When experts are not enough. Quantitative Science Studies **1**(1), 396–413 (2020)
39. Yang, B., Yih, W.t., He, X., Gao, J., Deng, L.: Embedding entities and relations for learning and inference in knowledge bases. arXiv preprint arXiv:1412.6575 (2014)
40. Yao, L., Zhang, Y., Wei, B., Jin, Z., Zhang, R., Zhang, Y., Chen, Q.: Incorporating knowledge graph embeddings into topic modeling. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)
41. Zhang, S., Tay, Y., Yao, L., Liu, Q.: Quaternion knowledge graph embedding. arXiv preprint arXiv:1904.10281 (2019)
42. Zhou, X., Zhu, Q., Liu, P., Guo, L.: Learning knowledge embeddings by combining limit-based scoring loss. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. pp. 1009–1018. ACM (2017)