IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 19, 2021, accepted August 5, 2021, date of publication August 16, 2021, date of current version August 26, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3105183

Link Prediction of Weighted Triples for Knowledge Graph Completion Within the Scholarly Domain

MOJTABA NAYYERI[1,2], GÖKCE MÜGE CIL[1], SAHAR VAHDATI[2], FRANCESCO OSBORNE[3], ANDREY KRAVCHENKO[4], SIMONE ANGIONI[5], ANGELO SALATINO[3], DIEGO REFORGIATO RECUPERO[5], ENRICO MOTTA[3], AND JENS LEHMANN[1,6]

[1]SDA Research Group, University of Bonn, 53115 Bonn, Germany
[2]Nature-Inspired Machine Intelligence, Institute for Applied Informatics (InfAI), 01069 Dresden, Germany
[3]Knowledge Media Institute, The Open University, Milton Keynes MK7 6AA, U.K.
[4]Christ Church, University of Oxford, Oxford OX1 1DP, U.K.
[5]Department of Mathematics and Computer Science, University of Cagliari, 09124 Cagliari, Italy
[6]Fraunhofer IAIS, 53757 Dresden, Germany

Corresponding author: Angelo Salatino (angelo.salatino@open.ac.uk)

This work was supported in part by the project Speaker under Grant BMWi FKZ 01MK20011A, in part by the project Joseph (Fraunhofer Zukunftsstiftung)–Cleopatra under Grant GA 812997, in part by the project Excellence Clusters ML2R under Grant BmBF FKZ 01 15 18038 A/B/C, in part by the project MLwin under Grant 01IS18050, in part by ScaDS.AI under Grant IS18026A-F, in part by the project Tailor under Grant EU GA 952215, and in part by the project H2020-EU PLATOON under Grant 872592.

ABSTRACT Knowledge graphs (KGs) are widely used for modeling scholarly communication, performing scientometric analyses, 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), leading to a reduced scope and quality of the resulting analyses. 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 of a certain triple. In this paper, we propose the *Weighted Triple Loss*, a new loss function for KGE models that takes full advantage of the additional numerical weights on facts and it is even tolerant to incorrect weights. 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 analyses and providing better support to systems for monitoring and predicting research dynamics.

INDEX TERMS Scholarly data, knowledge graphs, knowledge graph embeddings, loss functions, link prediction, scholarly communication, science of science.

I. 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 the scientific process [12]. 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.

The associate editor coordinating the review of this manuscript and approving it for publication was Ahmed A. Zaki Diab.

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 [17], [57]. 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) [19] and applying link prediction techniques [10] to this representation. Embedding models were successfully applied on KGs in different domains, including digital libraries [64], biomedical [23], and social media [50]. 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, or it further qualifies the information in the triple. Such a representation has already been described, analyzed and verified through a formal declarative semantics [55], [56]. 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.

In the scholarly domain the uncertainty typically stems by automatic approaches for disambiguating actors in this space [20] (e.g., authors, organizations, countries) or classifying articles according to specific categories [16], [41], [58] (e.g., topics, technologies, industrial sectors) as well as the limited coverage of complementary knowledge bases, such as GRID[1] (Global Research Identifier Database) and ORCID[2] (Open Researcher and Contributor ID). Since most of the existing KGE models can only handle triples that are either true or false, they perform quite poorly on KGs that contain weighted triples. There has been limited work on KGEs able to consider weighted triples. The main solution in this space is the Uncertain Knowledge Graph Embeddings (UKGE) [9], which however cannot properly handle erroneous or approximated weights in the graph. This is the situation, common in case of data incompleteness, in which the weights are potentially inaccurate.

In this paper, we propose the *Weighted Triple Loss*, a new loss function for KGE models that can effectively incorporate the numerical weights on facts and it is tolerant to incorrect or approximated weights. This loss is very general and can be used with different interaction models, e.g., DistMult [63], TransE [4], ComplEx [54]. We also introduce the *Weighted Rule Loss*, a loss function that extends the Rule Loss [34] in order to work with weighted triples. This solution exploits a set of automatically extracted logical rules to further improve performance.

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[3] in the field of Computer Science. AIDA classifies these documents according to the research topics from the Computer Science Ontology (CSO)[4] [42] and to the authors' affiliation types from the Global Research Identifier Database (GRID) (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 [40]. 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 [27], researcher migrations [29], collaboration patterns between academia and industry [2], and many others.

More in details, our main contributions are:

- The Weighted Triple Loss, a loss function for weighted triples which is agnostic with respect to their meaning and tolerant to incorrect weights.
- The *Weighted Rule Loss*, a second loss function for weighted triples that takes advantage of a set of automatically extracted logical rules.
- *AIDA35k*,[5] 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 II, we review the literature on current embedding models for data completion and scholarly knowledge graphs. In Section III, we present a motivating scenario involving the completion of the AIDA knowledge graph. In Section IV, we describe the architecture of the new optimization technique. Section V reports the evaluation of the model versus alternative

[1]GRID - https://www.grid.ac/
[2]ORCID - https://orcid.org/
[3]Dimensions - https://www.dimensions.ai/
[4]CSO - https://cso.kmi.open.ac.uk/
[5]AIDA35k - http://aida.kmi.open.ac.uk/aida35k/

solutions. Finally, in Section VI we summarise the main conclusions and outline future directions of research.

II. PRELIMINARIES AND RELATED WORK

In this section, we provide the background for knowledge graph embedding models (Section II-A) and then review the related work. Specifically, in Section II-B we give an overview of state-of-the-art KGE models that will be used in the evaluation of our work. In Section II-C we present some alternative methods for link prediction. In Section II-D we illustrate loss functions for KGEs. Finally, in Section II-E we discuss the knowledge graphs covering the scholarly domain.

A. KNOWLEDGE GRAPH EMBEDDING MODELS

A KGE model includes several components: embeddings (e.g., vector, matrix, tensor), a score function, and a loss function.

1) TRAINING SAMPLES

Each KGE model requires a set of samples used for training. The training set should contain both positive and negative samples where positive means true triples and negative means false triples. However, each KG $\mathcal{T} = \{(h, r, t)\}$ used for training contains only positive samples, where negative samples are not given explicitly. Therefore, for each individual positive sample, a set of negative samples $\mathcal{N}_{h,r,t} = \{(h', r, t')\}$ is randomly generated. This is performed by corrupting either the head h or the tail t. While most of KGs contain triples of the form (h, r, t), recent works focus on learning over KGs facts in the form of $(h, r, t, w_{h,r,t})$ where $w_{h,r,t}$ represents the weight of the triple h, r, t (e.g., the degree of uncertainty for the triples).

2) EMBEDDINGS

For a given triple (h, r, t), a KGE model mainly aims at obtaining vector representations for entities (shown in bold **h, t**) and the relation r involved in each triple. The embedding space can be real $\mathcal{R}^d$, complex $\mathcal{C}^d$ [51], [54] and quaternion $\mathcal{H}^d$ [65], which are generalized in geometric algebra $\mathcal{G}^d$ [62].

3) SCORE FUNCTION

A score function $f(h, r, t)$ takes the embeddings of a triple $(\mathbf{h}, \mathbf{r}, \mathbf{t})$ as input and returns a value reflecting its plausibility in the context of the KG. It is typically used by link prediction methods for assessing new candidate triples. Most modern KGE models either use distance functions (e.g., TransE [4], Trans4E [30], RotatE [51]) or inner product functions based on semantic matching (e.g., DistMult [63], ComplEx [54], QuatE [65], RESCAL [36], 5*E [31]) to compute scores of triples. The embeddings and consequently the scores of triples depend on the optimization of the loss function, which is introduced below.

4) LOSS FUNCTION

Typically KGE models employ loss functions that can be applied only on positive and negative triples, such as Margin Ranking Loss (MRL) [4], negative likelihood of logistic model [54], self-adversarial loss [51], Soft Mariginal [32] and full multiclass log loss [22]. An exception is RESCAL [36] 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 [9] combines the MSE loss with a rule-based loss in order to improve the ability to learn from weighted triples. In Section II-D we review these loss functions in detail.

B. OVERVIEW OF KGE MODELS

In this section, we provide a summary of the state-of-the-art KGE models that are also considered in the evaluation of this work. We consider two main classes of approaches: 1) distance-based models, which use distance functions (e.g. $L2$ norm) for score computation, and 2) semantic matching-based models, which use inner product.

1) DISTANCE-BASED KGE MODELS

TransE [4] is one of the primary translation-based KGE models and it is still considered a very competitive approach due to its simplicity and high performance. The score function of this model is:

$$f_r(h, t) = -\|h + r - t\|. \tag{1}$$

With this simple score function, the TransE model is mostly used as a baseline and outperforms many of the recent complex models.

RotatE [51] applies a rotation-based mechanism for transforming the head entity to the tail entity via a relation specific transformation. It uses a complex space for embedding the entities and the relations and its score function is:

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

where $\circ$ is an element-wise product. This model is currently one of the top performing KGE models for link prediction.

2) SEMANTIC MATCHING-BASED KGE MODELS

ComplEx [54] uses similarity of latent representations for scoring the positive and negative triples. The name of this model also represents the space in which it is designed, complex space. The underlying scoring function is: $f(h, t) = \Re(\mathbf{h}^T \operatorname{diag}(\mathbf{r})\, \bar{\mathbf{t}})$ where $\bar{\mathbf{t}}$ is the conjugate of the vector $\mathbf{t}$ and $\Re$ returns the real part of the complex number. It uses a matrix represented as $\operatorname{diag}(\mathbf{r})$ where the values of $\mathbf{r}$ are the diagonal elements of the matrix and the non-diagonal elements are zero.

QuatE [65] uses a mapping of $\mathcal{E} \rightarrow \mathcal{H}^d$, 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 of the QuatE model is:

$$\phi(h, r, t) = \mathbf{h}' \cdot \mathbf{t} = \langle a'_h, a_t \rangle + \langle b'_h, b_t \rangle + \langle c'_h, c_t \rangle + \langle d'_h, d_t \rangle \tag{3}$$

where $\langle \cdot, \cdot \rangle$ is the inner product.

For the computation of $\mathbf{h}'$, the relation embedding $\mathbf{r} = p_r + q_r\mathbf{i} + u_r\mathbf{j} + v_r\mathbf{k}$ is normalized to a unit quaternion:

$$\mathbf{r}^{(n)} = \frac{r}{|r|} = \frac{p_r + q_r\mathbf{i} + u_r\mathbf{j} + v_r\mathbf{k}}{\sqrt{p_r^2 + q_r^2 + u_r^2 + v_r^2}}. \quad (4)$$

Furthermore, the Hamiltonian product (shown by $\otimes$) between $\mathbf{r}^{(n)}$ and $\mathbf{h} = a_h + b_h\mathbf{i} + c_h\mathbf{j} + d_h\mathbf{k}$ is computed as:

$$\begin{aligned}\mathbf{h}' = \mathbf{h} \otimes \mathbf{r}^{(n)} := &(a_h \circ p - b_h \circ q - c_h \circ u - d_h \circ v) \\ &+ (a_h \circ q + b_h \circ p + c_h \circ v - d_h \circ u)\,\mathbf{i} \\ &+ (a_h \circ u - b_h \circ v + c_h \circ p + d_h \circ q)\,\mathbf{j} \\ &+ (a_h \circ v + b_h \circ u - c_h \circ q + d_h \circ p)\,\mathbf{k} \quad (5)\end{aligned}$$

DistMult [63] extends the **RESCAL** [36] model. The score function of RESCAL is:

$$f_r(h, t) = h^T M_r t, \quad (6)$$

where M_r is a relation-specific matrix. DistMult improves RESCAL using a matrix multiplication for capturing the relational semantics. The score function of this model uses a pairwise interaction of the latent features:

$$f_r(h, t) = h^T diag(r) t. \quad (7)$$

C. ALTERNATIVE METHODS FOR LINK PREDICTION

Beside shallow embedding models (e.g., QuatE, ComplEx, TransE, and RotatE) that are classified as one of the dimensionality reduction-based, many other techniques could be considered from information theory, clustering, and perturbation [26]. Graph neural networks (GNNs) [61] are one of the main techniques for link prediction which compute the state of embeddings for a node according to the local neighborhood [66]. Such models provide a d-dimension vector for each node and compute embeddings based on local neighborhood. Generally, GNNs have high computation costs which are problematic in large scale knowledge graphs. Another family of approaches that gained attention recently is the few shot learning (FSL) [60]. This does not deal with the structure of the data but the quantity of the data. In contrast with other machine learning approaches that need massive data to do accurate prediction and analysis, FSL approaches take smaller dataset as input and provide high performance output.

In terms of weighted triples, there are a few works from the Semantic Web community [6], [7], [24], [53] where RDF format is specified for weighted triples. However, in these works, the weights are not considered as quadratic information that extends the triple representation to quadruple, but as part of the tail of a triple. In [8], the time factor and uncertainty of triples are represented as weighted triples. However, they are not using any embedding model of prediction but Markov Logic Networks.

Some other non-embedding link prediction methods, such as [25], [46], [47], are also able to take into consideration weighted triples. However, they can only process simple undirected graphs, since they do not support multiple relation types and self-loops. Therefore, they are not applicable to most KGs, which are typically multi-relational directed graphs allowing self-loops for some relation types.

D. LOSS FUNCTIONS FOR KGEs

In this paragraph we review several loss functions that are used by state-of-the-art KGE models.

1) MARGIN RANKING LOSS

The margin ranking loss (MRL) [4], 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 sample (h', r, t'). The negative samples are generated by replacing either the 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. MRL has been widely used for training TransE and its variants.

2) LIMIT-BASED SCORING LOSS

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 (e.g., TransE), such a limitation would prevent the model from fulfilling the translation in the vector space, resulting in poor performance [67]. The limit-based scoring loss [67] aims at avoiding this issue by including the boundary for the scores of positive triples in the margin ranking loss.

$$\begin{aligned}\mathcal{L}_{LSL} = \sum_{(h,r,t)\in\mathcal{T}} \sum_{(h',r,t')\in\mathcal{N}_{h,r,t}} &[f(h, r, t) + \gamma - f(h', r, t')]_+ \\ &+ \lambda[f(h, r, t) - \gamma_1]_+. \quad (8)\end{aligned}$$

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 (by γ_1) not to be very large. λ is a multiplier of the regularization term that determines the degree of importance of the term $[f(h, r, t) - \gamma_1]_+$ in the optimization. This loss function improved the performance of translation based embedding models (TransE, TransH, TransR etc).

3) SOFT MARGIN LOSS (SML)

The soft margin loss [33] aims at handling noisy negative samples. It adds slack variables ($\eta_{h,r,t}$) to negative samples optimization in order to mitigate the negative effect of false negative samples:

$$\begin{aligned}\mathcal{L}_{SML} = \sum_{(h,r,t)\in\mathcal{T}} \sum_{(h',r,t')\in\mathcal{N}_{h,r,t}} &\lambda\,\eta_{h,r,t}^2 + \lambda_+[f(h, r, t) - \gamma_1]_+ \\ &+ \lambda_-\,[\gamma_2 - f(h', r, t') - \eta_{h,r,t}]_+ \quad (9)\end{aligned}$$

where $\lambda, \lambda_+, \lambda_-$ are regularization weights used as hyperparameters. This loss function has been used for training TransE and RotatE.

4) SlidE LOSS

The length of the margin is an important factor which affects the performance of KGE models. In Limit-based Scoring Loss and Soft Margin Loss the length of the margin is determined by setting two hyper-parameters by a trial and error process. The SlidE loss [35] addresses this issue by determining the center of the margin and automatically adjusting the length of the margin by means of a trainable variable (η). The formulation of the SlidE loss is as follows:

$$\mathcal{L}_{SlidE^+} = \lambda e^{-\sigma\eta^2} + \lambda_+ [f(h,r,t) - \gamma + \eta]_+ + \lambda_- [-f(h',r,t') + \gamma + \eta]_+, \quad (10)$$

where γ is the center of margin, 2η is the length of margin, and σ is the variance of the Gaussian function that affects the margin.

5) SELF ADVERSARIAL LOSS (SAL)

Self-Adversarial Loss [51] obtained state-of-the-art performances on distance based KGE models. Its formulation is:

$$\mathcal{L} = -\sum_{(h,r,t)\in\mathcal{T}} \log\sigma(\gamma - f(h,r,t)) + \sum_{(h',r,t')\in\mathcal{N}_{h,r,t}} p(h',r,t')\log\sigma(f(h',r,t') - \gamma), \quad (11)$$

where $p(h',r,t') = \frac{exp(\alpha f(h',r,t'))}{\sum exp(\alpha f(h',r,t'))}$, α is adversarial temperature which represents the extent of attention on the score of negative samples used for random sampling. The loss assigns higher weights for negative samples with high values to reduce their scores as much as possible. σ is Sigmoid function.

6) NEGATIVE LOG LIKELIHOOD LOSS (NLL)

The negative likelihood [54] of the logistic model with regularization is:

$$\mathcal{L}_{NLL} = \sum_{(h,r,t)\in\mathcal{T}\cup\mathcal{N}} \lambda\ \log(1 + exp(-y_{h,r,t} f(h,r,t))) + \lambda||\theta||^2, \quad (12)$$

where θ is used to represent all adjustable parameters for simplicity purpose and $y_{h,r,t}$ is the label of triples (positive triples are labeled with 1 and negative triples are labeled with -1).

7) FULL MULTICLASS LOG LOSS (FMLL)

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 performances by using the full multiclass log-loss [22]. The loss applies full negative sampling and is defined as:

$$\mathcal{L}_{FMLL} = \sum_{(h,r,t)\in\mathcal{T}} l(f(h,r,t))), \quad (13)$$

where $l(f(h,r,t))) = l^1(f(h,r,t)) + l^2(f(h,r,t))$, $l^1(f(h,r,t)) = -f(h,r,t) + log(\sum_{t'} exp(f(h,r,t')))$, $l^2(f(h,r,t)) = -f(h,r,t) + log(\sum_{h'} exp(f(h',r,t)))$, where *log* and *exp* are logarithmic and exponential functions, respectively. The loss gives big (small) scores for positive (negative) triples.

8) UKGE LOSS

The previously discussed loss functions are suitable for learning over triples which are either positive or negative and cannot handle weighted triples.

Recently, the UKGE model used the MSE loss (already adopted by the RESCAL model) together with probabilistic soft logic loss, for training KGs with uncertain triples i.e., triples associated with a weight that reflects the confidence of their correctness [9]. This loss is model independent and it is formulated as:

$$\mathcal{L}_{UKGE} = \sum_{(h,r,t)\in\mathcal{T}_w\cup\mathcal{N}} |f(h,r,t) - w_{h,r,t}|^2 + \sum_{(h,r,t)}\sum_{g} |\psi_g(f(h,r,t))|^2, \quad (14)$$

where $w_{h,r,t}$ is the weight of a triple (h,r,t) and $\mathcal{T}_w$ is the set of all the weighted triples. g refers to a rule and ψ_g is the weighted distance of the rule g obtained by probabilistic soft logic.

An important limitation of this loss is that it constrains the KGE model to learn scores that are very close to the input ones. This can be problematic when dealing with approximated or incorrect values that are then incorporated in the model without any correction. We will discuss this issue further in Section IV-A.

E. 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) [59], 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[6] is a dataset of about 185M publications released by Semantic Scholar, an academic search engine provided by the Allen Institute for Artificial Intelligence. One more knowledge graph is the OpenCitations Corpus [39], that 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)[7] is a large knowledge graph integrating 208M papers from MAG and 172M from AMiner.

[6]ORC - http://s2-public-api-prod.us-west-2.elasticbeanstalk.com/corpus/
[7]OAG - https://www.openacademic.ai/oag/

All these resources suffer from data incompleteness to different degrees. For instance, it is still challenging to identify and disambiguate affiliations. This hinders our ability to categorize the articles according to their affiliation type or country [27]. Similarly, references are usually incomplete, and the citation count of the same paper tends to vary dramatically on different datasets [39].

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 [37], encouraging the Semantic Publishing paradigm [48], implementing systems for managing nano-publications [14], [21] and micropublications [44], and developing a variety of ontologies to describe scholarly data, e.g., SWRC,[8] BIBO,[9] BiDO,[10] SPAR [38],[11] CSO [43]. A recent project is the Open Research Knowledge Graph (ORKG) [18], which aims at describing research papers in a structured manner to make them easier to find and compare. Similarly, the Artificial Intelligence Knowledge Graph (AI-KG)[12] 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.

III. 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)[13] [59] 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) [42], 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 [40].

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 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 from a company. Overall this data show that these organization types are very different in terms 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.

For instance, in Computer Science only 8.1% of the articles involving authors from Universities (Education) include at least one collaborator from the other categories[14] (e.g., Company, Government). Conversely, authors from companies collaborate with at least another category (mostly Education) in 14.6% of the cases. This number raises to 46.1% for Government Institutions, 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%).

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 associated with 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

[8]SWRC - http://ontoware.org/swrc
[9]BIBO - http://bibliontology.com
[10]BiDO - http://purl.org/spar/bido
[11]SPAR - http://www.sparontologies.net/
[12]AI-KG - http://scholkg.kmi.open.ac.uk/
[13]MAG - https://academic.microsoft.com/

[14]This percentage is computed as the difference between the number of articles in Computer Science with at least an author from Education (*Computer Science (>0)* in Table 1, 3,969,096) and the number of articles in Computer Science with only authors from Education (*Computer Science (=1.0)*, 3,648,629).

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

associated with 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 with 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.

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 functions, we apply them on a subset of the AIDA knowledge graph which we named AIDA35k, a new dataset including 35K entities from AIDA associated with 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).

IV. OPTIMISING KGE MODELS FOR WEIGHTED TRIPLES

In this section we propose two loss functions: Weighted Triple Loss and Rule Loss for Weighted Triples. These loss functions optimise the weighted triples of the form $(h, r, t, w_{h,r,t})$ where h and t are the 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).

A. 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 triple 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 (WTL) presented in this paper is that in many practical cases the weight $w_{h,r,t}$ is estimated on the basis of potentially incomplete data and possibly biased computational methods. For instance, the weights associated with the organization types in AIDA depend on many factors such as the coverage of the GRID database in a particular moment in time and the performance of the disambiguation approaches applied by MAG. Therefore, these weights are typically approximations and some of them may be simply incorrect. This limitation needs to be taken into account during the learning phase. Therefore WTL allows the model to learn the score $f(h, r, t)$ of a triple (h, r, t) in the range:

$$w_{h,r,t} - \eta^{-2}_{h,r,t} \leq f(h, r, t) \leq w_{h,r,t} + \eta^{+2}_{h,r,t}, \tag{15}$$

where $\eta^{-2}_{h,r,t}$ and $\eta^{+2}_{h,r,t}$ 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 $w_{h,r,t} - \eta^{-2}_{h,r,t}$ and $w_{h,r,t} + \eta^{+2}_{h,r,t}$. In order to optimize the embedding vectors of entities and relations as well as adjusting $\eta^{-2}_{h,r,t}$ and $\eta^{+2}_{h,r,t}$, the following optimization framework is proposed:

$$\begin{cases} \min_\theta \sum_{(h,r,t,w_{h,r,t}) \in \{\mathcal{T}_w \cup \mathcal{N}\}} \lambda_1 \eta^{-2}_{h,r,t} + \lambda_2 \eta^{+2}_{h,r,t} + \lambda_3 \mathcal{L}, \\ \text{s.t. } w_{h,r,t} - \eta^{-2}_{h,r,t} \leq f(h, r, t) \leq w_{h,r,t} + \eta^{+2}_{h,r,t}, \end{cases} \tag{16}$$

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 the regularization term over the embeddings of entities and relations, θ contains all the adjustable parameters including the embeddings of entities, relations and $\eta^{-2}_{h,r,t}, \eta^{+2}_{h,r,t}$ i.e. $\theta = \{(\mathbf{h}, \mathbf{r}, \mathbf{t}) \cup \{\eta^{-2}_{h,r,t}, \eta^{+2}_{h,r,t}\} | (h, r, t) \in \mathcal{T}\}$. $\mathcal{L}$ is the regularization over the entities and relations embeddings i.e. $\mathcal{L} = \mathbf{E^2} + \mathbf{R^2}$. $\mathbf{E}$ and $\mathbf{R}$ are the embeddings of all the 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 the corrupted samples, we set their weights w to zero. We indicate the set of all corrupted samples by $\mathcal{N}$.

This solution results in a high tolerance to incorrect weights. Indeed, the UKGE [9] loss forces the KGE model to

FIGURE 1. RDF Schema for articles in AIDA.

learn scores that are very close to the input weights, therefore incorrect values are preserved and incorporated in the model. Conversely, WTL allows for more flexibility in the learning process, by allowing for a wider range of scores to be learnt and as a result the incorrect weights can be corrected by using contextual information from other triples.

B. WEIGHTED RULE LOSS

1) EXTRACTION OF RULES

In order to include additional logical rules as complementary knowledge, we used the AMIE rule extractor [13], 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 AMIE.

2) 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.* [34]. For a given rule of the form $rule: q_1 \wedge q_2 \wedge, \ldots, \wedge q_n \rightarrow q_{n+1}$ where $q_i,\ i = 1, \ldots, n+1$ 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.

$$\mathcal{R} = \max(w_{q_1} *, \ldots, * w_{q_n} - f(q_{n+1}), 0), \tag{17}$$

where w_{q_i} is the weight of the weighted triples $q_i, i = 1, \ldots, n$ after grounding of entities (replacing the 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

$$\begin{cases} \min_\theta \sum_{(h,r,t,w_{h,r,t}) \in \{\mathcal{T}_w \cup \mathcal{N}\}} \lambda_1 \eta_{h,r,t}^{-2} + \lambda_2 \eta_{h,r,t}^{+2} \\ +\lambda_3 \mathcal{L} + \lambda_4 \sum_{i=1}^{l} \mathcal{R}_i, \\ \text{s.t. } w_{h,r,t} - \eta_{h,r,t}^{-2} \leq f(h, r, t) \leq w_{h,r,t} + \eta_{h,r,t}^{+2}, \end{cases} \tag{18}$$

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

$$\begin{cases} \min_\theta \sum_{(h,r,t,w_{h,r,t}) \in \{\mathcal{T} \cup \mathcal{T}'_w \cup \mathcal{N}\}} \lambda_1 \eta_{h,r,t}^{-2} + \lambda_2 \eta_{h,r,t}^{+2} \\ +\lambda_3 \mathcal{L}, \text{ s.t.} \\ w_{h,r,t} - \eta_{h,r,t}^{-2} \leq f(h, r, t) \leq w_{h,r,t} + \eta_{h,r,t}^{+2}. \end{cases} \tag{19}$$

V. 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 *et al.* [9] (see Section II-D.h), iii)

TABLE 2. Representation of some example rules.

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

TABLE 3. Performance of approaches on the three benchmarks (NL27k, PPI5k, CN15k). In bold the best results.

	PPI5K					*CN15k*					*NL27k*				
	MSE	**MAE**	**F1**	**Acc.**	**AUC**	**MSE**	**MAE**	**F1**	**Acc.**	**AUC**	**MSE**	**MAE**	**F1**	**Acc.**	**AUC**
UKGE_rect	**.003**	.028	**.969**	**.997**	**.986**	.167	.330	.269	.904	.598	.027	.071	.931	.957	.940
UKGE_rect_psl	.003	.028	.968	.997	**.986**	.166	.329	.268	.903	.597	.026	.072	.931	.957	.939
WGE_rect	.004	**.022**	.912	.992	.985	**.148**	**.307**	**.304**	**.902**	**.614**	**.019**	**.065**	.901	.934	.925
UKGE_logi	.005	.035	.967	**.997**	.979	.281	.448	.258	.901	.591	.055	.111	.904	.942	.917
UKGE_logi_psl	.005	.035	.966	**.997**	.982	.279	.446	.255	.901	.590	.056	.114	.802	.864	.859
WGE_logi	.009	.049	.921	.993	**.986**	.223	.388	.279	.861	.603	.032	.076	**.937**	**.960**	**.941**
TransE	-	-	.832	.985	-	-	-	.234	.679	-	-	-	.651	.534	-
DistMult	-	-	.869	.979	-	-	-	.279	.711	-	-	-	.721	.701	-
Complex	-	-	.832	.989	-	-	-	.189	.732	-	-	-	.633	.534	-

TABLE 4. Performance of approaches on AIDA35k. In bold the best results. The reported time is in seconds.

	AIDA35k					
	MSE	**MAE**	**F1**	**Acc.**	**AUC**	**Time**
UKGE_rect	.215	.299	.698	.795	.769	5.8
UKGE_rect_rules	.204	.290	.719	.803	.777	6.9
WGE_rect	.143	.179	.842	.865	.852	6.6
WGE_rect_rules	.135	.170	**.847**	**.871**	**.864**	7.4
UKGE_logi	.234	.341	.699	.731	.726	5.5
UKGE_logi_rules	.209	.317	.720	.783	.752	6.8
WGE_logi	**.087**	.129	.750	.743	.665	6.2
WGE_logi_rules	.097	**.128**	.808	.820	.779	7.5
TransE	-	-	.674	.642	.670	5.8
DistMult	-	-	.741	.642	.766	5.4
ComplEx	-	-	.748	.778	.775	**5.0**

TABLE 5. Statistical information about the datasets. Avg(s) and Std(s) are the average and standard deviation of the scores.

Dataset	Entities	Relations	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

DistMult [63], iv) *TransE* [4], and v) *ComplEx* [54] on several datasets.

In addition to AIDA35k, which was introduced in Section 3, we used three other datasets that include weighted triples: CN15k, NL27k, and PPI5k. These were used in the evaluation of UKGE [9], which is one of the baselines. CN15k is a subgraph of ConceptNet [49] that covers 15,000 entities and 241,158 uncertain relation facts in English. NL27k was obtained from NELL [28], an uncertain KG extracted from webpages containing information about cities, companies, emotions and sports teams. It covers 27,221 entities, 404 relations, and 175,412 uncertain relation facts. Finally, PPI5k was extracted from the Protein-Protein Interaction Knowledge Base STRING [52] and contains 271,666 uncertain relation facts for 4,999 proteins and 7 interactions.

The evaluation data are available at http://aida.kmi.open.ac.uk/aida35k/.

We considered two different versions of the WGE and UKGE models using two score functions, respectively the logistic function (**WGE_logi** and **UKGE_logi**) and the bounded rectifier (**WGE_rect** and **UKGE_rect**) [9]. These are defined as follows:

- **Logistic Function:** $\Phi(x) = \frac{1}{1+e^{-(wx+b)}}$
- **Bounded Rectifier:** $\Phi(x) = min(max(wx+b, 0), 1)$.

In addition, since the original article about UKGE also presented an alternative version that injects probabilistic soft logic (PSL) rules for deriving weights between 0 and 1 for unobserved triples, we also considered other two alternative versions of UKGE that make use of PSL (**UKGE_rect_psl** and **UKGE_logi_psl**). However, these models could only be used for the three datasets released with the original paper about UKGE (PP15k, CN15k, NL27k) [9], since the article does not give enough details to reproduce these models on a new dataset.

On AIDA35k, we further tested two versions of WGE (**WGE_rect_rules** and **WGE_logi_rules**) and two versions of UKGE (**UKGE_rect_rules** and **UKGE_logi_rules**) that use the Weighted Rules Loss as defined in Section IV-B. In order to extract the rules from AIDA35k we used AMIE+ [13]. This results in initial set of about 40 rules from which we filtered strong rules only (18) and categorized them to produce the corresponding groundings (around 126k).

TABLE 6. 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	0.049	**0.149**	**0.848**	**0.882**	0.057	**0.177**	0.696	0.790
WGE_logi	**0.045**	0.152	0.694	0.702	0.136	0.243	0.613	0.689
WGE_rect_rules	0.060	0.175	0.833	0.864	**0.053**	0.183	**0.713**	**0.804**
WGE_logi_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

When we run AMIE+ on the other three datasets, it produced an unfeasible number of rules (more than 10k rules) to be integrated in the model. Therefore, we limited the evaluation of the Weighted Rules Loss to the AIDA35k dataset.

A. EXPERIMENTAL SETUP

1) ENVIRONMENT

We implemented our model WGE using Python 3.7 and PyTorch (version 1.7.1) library. We used the Sklearn library (version 0.22) for implementing the evaluation metrics. Furthermore we adopted Adam as optimizer for training our model. We employed AMIE+ [13][15] to automatically extract logical rules from the KG and ran the code on Google Colab servers using a Tesla K80 GPU and 13 GB of RAM.

The code of our approach can be freely accessed at https://github.com/gokcemuge/WeightedGraphEmbedding, while the data used for training and evaluation are publicly available at http://aida.kmi.open.ac.uk/aida35k/.

2) 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 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.

3) METRICS AND HYPERPARAMETERS

We adopted the Mean Square Error (MSE), the Mean Absolute Error (MAE), the Area Under Curve (AUC) [45] and the F1 measure as evaluation metrics. We also evaluated the time complexity of our approach with the granularity of seconds per epochs. Since the space is limited and they are standard metrics used by most works in this field [15], they will not be described in this paper.

The setup 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 $\{0.1, 0.5, 1.0\}$.

[15]https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/amie/

B. RESULTS

Table 3 reports the performance of the approaches on the PPI5k, CN15k, and NL27. On the NL27k dataset, the rectifier version of WGE (*WGE_rect*) outperforms the other approaches in MSE and MAE, while the logistic version (*WGE_logi*) achieves the best results in F1, AUC 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. On CN15k the rectified WGE (*WGE_rect*) obtains the highest performance in all the metrics.

Table 4 reports the performance and running time on the AIDA35k dataset. WGE_rect_rules obtains the best results in terms of AUC (0.864) , F1 (0.847) and accuracy (0.871), while WGE_logi_rules yields the best MAE and WGE_logi the best MSE. While our model outperforms the other competitors, the running time (reported in seconds per epochs) of our model is close to that of other models.

Table 6 zooms on two specific relations from AIDA35k: *hasGridType* and *hasCountry*. For *hasGridType*, *WGE_rect* achieves the best results for MAE, F1, and accuracy. For *hasCountry*, the rectified version of WGE with rules (*WGE_rect_rules*) outperforms all the other models in MSE, F1, and accuracy. This suggest that incorporating the Weighted Rule Loss can enhance significantly the performance, especially for types of certain relations.

Overall, WGE, our solution based on WTL, outperformed UKGE on AIDA35k, CN15k, and NL27k and obtainsed competitive results on PP15k. This seems to be due to the ability of WTL of tolerating better incorrect weights. WGE also outperformed by a large margin the other models based on loss functions that do not handle weighted triples. In particular, the difference in terms of F1 score and accuracy between the standard DistMult model and the DistMult interaction model with the best variant of our proposed loss function is higher than 10% on average across all datasets. This empirically confirms our hypothesis that there is a substantial benefit

in using triple weight information, if available, in the loss function.

VI. CONCLUSION AND FUTURE WORK

In this paper we proposed the *Weighted Triple Loss* (WTL), a new loss function for KGE models that can effectively handle weighted triples and is tolerant to incorrect or approximated weights. We also introduced the *Weighted Rule Loss*, a loss function that extends the Rule Loss in order to work with weighted triples.

In order to test these solutions, we developed the *Weighted Graph Embedding* (WGE), a new KGE model which uses the interaction model of DistMult with the two loss functions.

The evaluation showed that this approach outperforms all the baselines (UKGE, TransE, Distmult, and ComplEx) and achieves higher result than baseline on AIDA35k (metadata of research articles), NL27k (data from web pages), and CN15k (concepts from ConceptNet). It also obtains competitive results on PPI5k (proteins from STRING).

WGE was originally designed to address the real world scenario of completing the AIDA Knowledge Graph, in order to enable more comprehensive quantitative analysis of science about geopolitical factors [27] and the flow of knowledge between different types of organizations [2] (e.g., university, industry, non-profit). However, the loss functions presented in this paper are general solutions that can be used in many different domain in order to take into account the weighted triples. They can also can support different interaction models, such as DistMult [63], TransE [4], ComplEx [54].

The approach presented in this paper opens up several interesting directions of work. First, we aim to apply WGE on other KGs in this space for improving their coverage of the research landscape. Specifically, we plan to run it on recent KGs describing scientific concepts (e.g., tasks, methods, materials) and their relationships, such as AI-KG [11] and ORKG [18], where the numerical weights could represent the consensus of the research community on the relevant statements. We also plan to apply model selections techniques in order to investigate the best set of parameters and evaluation methods in this space. Finally, we would also like to apply our approach to a range of KGs in other domains for investigating how the results and performances might be affected by the underlying domain.

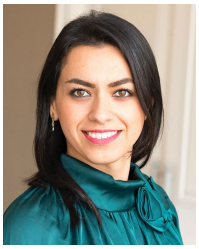
REFERENCES

[1] S. Angioni, A. A. Salatino, F. Osborne, D. R. Recupero, and E. Motta, "Integrating knowledge graphs for analysing academia and industry dynamics," in *Proc. ADBIS, TPDL EDA Common Workshops Doctoral Consortium*, vol. 1260, L. Bellatreche *et al.*, Eds. Cham, Switzerland: Springer, 2020, doi: 10.1007/978-3-030-55814-7_18.
[2] S. N. Ankrah and O. Al-Tabbaa, "Universities-industry collaboration: A systematic review," *SSRN Electron. J.*, vol. 31, no. 3, pp. 387–408, 2015.
[3] M. Awad and R. Khanna, "Support vector regression," in *Efficient Learning Machines*. Berkeley, CA, USA: Apress, 2015, doi: 10.1007/978-1-4302-5990-9_4.
[4] A. Bordes, N. Usunier, A. Garcia-Durán, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in *Proc. 26th Int. Conf. Neural Inf. Process. Syst. (NIPS)*, vol. 2. Red Hook, NY, USA: Curran Associates, 2013, pp. 2787–2795.
[5] J. Brainard, "Scientists are drowning in COVID-19 papers. Can new tools keep them afloat," *Science*, vol. 13, no. 10, p. 1126, 2020.
[6] J. P. Cedeño and K. S. Candan, "R2DF framework for ranked path queries over weighted RDF graphs," in *Proc. Int. Conf. Web Intell., Mining Semantics (WIMS)*, 2011, pp. 1–12.
[7] J. P. Cedeno, "A framework for top-K queries over weighted RDF graphs," Ph.D. dissertation, Arizona State Univ., Tucson, AZ, USA, 2010. [Online]. Available: https://repository.asu.edu/items/8620
[8] M. W. Chekol, G. Pirrò, J. Schoenfisch, and H. Stuckenschmidt, "Marrying uncertainty and time in knowledge graphs," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 88–94.
[9] X. Chen, M. Chen, W. Shi, Y. Sun, and C. Zaniolo, "Embedding uncertain knowledge graphs," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 3363–3370.
[10] Y. Dai, S. Wang, N. N. Xiong, and W. Guo, "A survey on knowledge graph embedding: Approaches, applications and benchmarks," *Electronics*, vol. 9, no. 5, p. 750, May 2020.
[11] D. Dessì, F. Osborne, D. R. Recupero, D. Buscaldi, E. Motta, and H. Sack, "AI-KG: An automatically generated knowledge graph of artificial intelligence," in *The Semantic Web–(ISWC)*, J. Z. Pan, V. Tamma, C. d'Amato, K. Janowicz, B. Fu, A. Polleres, O. Seneviratne, and L. Kagal, Eds. Cham, Switzerland: Springer, 2020, pp. 127–143.
[12] S. Fortunato, C. T. Bergstrom, K. Börner, J. A. Evans, D. Helbing, S. Milojević, A. M. Petersen, F. Radicchi, R. Sinatra, and B. Uzzi, "Science of science," *Science*, vol. 359, no. 6379, pp. eaao0185, 2018.
[13] L. Galárraga, C. Teflioudi, K. Hose, and F. M. Suchanek, "Fast rule mining in ontological knowledge bases with AMIE+," *VLDB J.*, vol. 24, no. 6, pp. 707–730, 2015.
[14] P. Groth, A. Gibson, and J. Velterop, "The anatomy of a nanopublication," *Inf. Services Use*, vol. 30, nos. 1–2, pp. 51–56, Sep. 2010.
[15] A. Gunawardana and G. Shani, "A survey of accuracy evaluation metrics of recommendation tasks," *J. Mach. Learn. Res.*, vol. 10, pp. 2935–2962, Dec. 2009.
[16] V. Henk, S. Vahdati, M. Nayyeri, M. Ali, H. S. Yazdi, and J. Lehmann, "Metaresearch recommendations using knowledge graph embeddings," in *Proc. RecNLP Workshop AAAI Conf.*, 2019.
[17] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. de Melo, C. Gutierrez, J. Emilio Labra Gayo, S. Kirrane, S. Neumaier, A. Polleres, R. Navigli, A.-C. Ngonga Ngomo, S. M. Rashid, A. Rula, L. Schmelzeisen, J. Sequeda, S. Staab, and A. Zimmermann, "Knowledge graphs," 2020, *arXiv:2003.02320*. [Online]. Available: http://arxiv.org/abs/2003.02320
[18] M. Y. Jaradeh, A. Oelen, K. E. Farfar, M. Prinz, J. D'Souza, G. Kismihók, M. Stocker, and S. Auer, "Open research knowledge graph: Next generation infrastructure for semantic scholarly knowledge," in *Proc. 10th Int. Conf. Knowl. Capture*, 2019, pp. 243–246.
[19] S. Ji, S. Pan, E. Cambria, P. Marttinen, and P. S. Yu, "A survey on knowledge graphs: Representation, acquisition and applications," 2020, *arXiv:2002.00388*. [Online]. Available: http://arxiv.org/abs/2002.00388
[20] J. Kim, "Evaluating author name disambiguation for digital libraries: A case of DBLP," *Scientometrics*, vol. 116, no. 3, pp. 1867–1886, Sep. 2018.
[21] T. Kuhn, C. Chichester, M. Krauthammer, N. Queralt-Rosinach, R. Verborgh, G. Giannakopoulos, A.-C. Ngonga Ngomo, R. Viglianti, and M. Dumontier, "Decentralized provenance-aware publishing with nanopublications," *PeerJ Comput. Sci.*, vol. 2, p. e78, Aug. 2016.
[22] T. Lacroix, N. Usunier, and G. Obozinski, "Canonical tensor decomposition for knowledge base completion," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2863–2872.
[23] L. Li, P. Wang, Y. Wang, J. Jiang, B. Tang, J. Yan, S. Wang, and Y. Liu, "A method to learn embedding of a probabilistic medical knowledge graph: Algorithm development," 2019, *arXiv:1909.00672*. [Online]. Available: http://arxiv.org/abs/1909.00672
[24] S. Liu, J. P. Cedeno, K. S. Candan, M. L. Sapino, S. Huang, and X. Li, "R2DB: A system for querying and visualizing weighted RDF graphs," in *Proc. IEEE 28th Int. Conf. Data Eng.*, Apr. 2012, pp. 1313–1316.
[25] L. Lü and T. Zhou, "Link prediction in weighted networks: The role of weak ties," *Europhys. Lett.*, vol. 89, no. 1, p. 18001, Jan. 2010.
[26] L. Lü and T. Zhou, "Link prediction in complex networks: A survey," *Phys. A, Statist. Mech. Appl.*, vol. 390, no. 6, pp. 1150–1170, 2011.
[27] A. Mannocci, F. Osborne, and E. Motta, "Geographical trends in academic conferences: An analysis of authors' affiliations," *Data Sci.*, vol. 2, nos. 1–2, pp. 181–203, 2019.
[28] T. Mitchell, W. Cohen, E. Hruschka, P. Talukdar, B. Yang, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel, and J. Krishnamurthy, "Never-ending learning," *Commun. ACM*, vol. 61, no. 5, pp. 103–115, 2018.
[29] H. F. Moed, M. Aisati, and A. Plume, "Studying scientific migration in scopus," *Scientometrics*, vol. 94, no. 3, pp. 929–942, Mar. 2013.
[30] M. Nayyeri, G. M. Cil, S. Vahdati, F. Osborne, M. Rahman, S. Angioni, A. Salatino, D. R. Recupero, N. Vassilyeva, E. Motta, and J. Lehmann, "Trans4E: Link prediction on scholarly knowledge graphs," *Neurocomputing*, 2021, doi: 10.1016/j.neucom.2021.02.100.

[31] M. Nayyeri, S. Vahdati, C. Aykul, and J. Lehmann, "5* knowledge graph embeddings with projective transformations," in *Proc. AAAI Conf. Artif. Intell.*, 2021.
[32] M. Nayyeri, S. Vahdati, J. Lehmann, and H. Shariat Yazdi, "Soft marginal TransE for scholarly knowledge graph completion," 2019, *arXiv:1904.12211*. [Online]. Available: http://arxiv.org/abs/1904.12211
[33] M. Nayyeri, S. Vahdati, X. Zhou, H. S. Yazdi, and J. Lehmann, "Embedding-based recommendations on scholarly knowledge graphs," in *The Semantic Web* (Lecture Notes in Computer Science), vol. 12123, A. Harth *et al.*, Eds. Cham, Switzerland: Springer, 2020, doi: 10.1007/978-3-030-49461-2_15.
[34] M. Nayyeri, C. Xu, J. Lehmann, and H. Shariat Yazdi, "LogicENN: A neural based knowledge graphs embedding model with logical rules," 2019, *arXiv:1908.07141*. [Online]. Available: http://arxiv.org/abs/1908.07141
[35] M. Nayyeri, X. Zhou, S. Vahdati, R. Izanloo, H. S. Yazdi, and J. Lehmann, "Let the margin SlidE$^{\pm}$ for knowledge graph embeddings via a correntropy objective function," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1–9.
[36] M. Nickel, V. Tresp, and H.-P. Kriegel, "A three-way model for collective learning on multi-relational data," in *Proc. ICML*, vol. 11, 2011, pp. 809–816.
[37] A. G. Nuzzolese, A. L. Gentile, V. Presutti, and A. Gangemi, "Semantic web conference ontology—A refactoring solution," in *The Semantic Web* (Lecture Notes in Computer Science), vol. 9989, H. Sack, G. Rizzo, N. Steinmetz, D. Mladenić, S. Auer, and C. Lange, Eds. Cham, Switzerland: Springer, 2016, doi: 10.1007/978-3-319-47602-5_18.
[38] S. Peroni and D. Shotton, "The SPAR ontologies," in *The Semantic Web—ISWC 2018* (Lecture Notes in Computer Science), vol. 11137, D. Vrandečić *et al.*, Eds. Cham, Switzerland: Springer, 2018, doi: 10.1007/978-3-030-00668-6_8.
[39] S. Peroni and D. Shotton, "OpenCitations, an infrastructure organization for open scholarship," *Quant. Sci. Stud.*, vol. 1, no. 1, pp. 428–444, Feb. 2020.
[40] A. Salatino, F. Osborne, and E. Motta, "Researchflow: Understanding the knowledge flow between academia and industry," in *Proc. 22nd Int. Conf. Knowl. Eng. Knowl. Manage.*, 2020, pp. 219–236.
[41] A. Angelo Salatino, F. Osborne, T. Thanapalasingam, and E. Motta, "The CSO classifier: Ontology-driven detection of research topics in scholarly articles," in *Digital Libraries for Open Knowledge*. Cham, Switzerland: Springer, 2019, pp. 296–311.
[42] A. A. Salatino, T. Thanapalasingam, A. Mannocci, A. Birukou, F. Osborne, and E. Motta, "The computer science ontology: A comprehensive automatically-generated taxonomy of research areas," *Data Intell.*, vol. 2, no. 3, pp. 379–416, Jul. 2020.
[43] A. A. Salatino, T. Thanapalasingam, A. Mannocci, F. Osborne, and E. Motta, "The computer science ontology: A large-scale taxonomy of research areas," in *Proc. Int. Semantic Web Conf.*, 2018, pp. 187–205.
[44] J. Schneider, P. Ciccarese, T. Clark, and R. D. Boyce, "Using the micropublications ontology and the open annotation data model to represent evidence within a drug-drug interaction knowledge base," in *Proc. 4th Int. Conf. Linked Sci.*, vol. 1282. Aachen, Germany: CEUR-WS.org, 2014, pp. 60–70.
[45] K.-K. Shang, M. Small, and W.-S. Yan, "Fitness networks for real world systems via modified preferential attachment," *Phys. A, Stat. Mech. Appl.*, vol. 474, pp. 49–60, May 2017.
[46] K.-K. Shang, M. Small, X.-K. Xu, and W.-S. Yan, "The role of direct links for link prediction in evolving networks," *Europhys. Lett.*, vol. 117, no. 2, p. 28002, Jan. 2017.
[47] K.-K. Shang, M. Small, D. Yin, T.-C. Li, and W. Yan, "The key to the weak-ties phenomenon," *EPL (Europhys. Lett.)*, vol. 127, no. 4, p. 48002, Sep. 2019.
[48] D. Shotton, "Semantic publishing: The coming revolution in scientific journal publishing," *Learned Publishing*, vol. 22, no. 2, pp. 85–94, Apr. 2009.
[49] R. Speer, J. Chin, and C. Havasi, "ConceptNet 5.5: An open multilingual graph of general knowledge," in *Proc. 31st AAAI Conf. Artif. Intell. (AAAI)*. AAAI Press, 2017, pp. 4444–4451.
[50] G. Stanovsky, D. Gruhl, and P. Mendes, "Recognizing mentions of adverse drug reaction in social media using knowledge-infused recurrent models," in *Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics*, vol. 1, 2017, pp. 142–151.
[51] Z. Sun, Z.-H. Deng, J.-Y. Nie, and J. Tang, "Rotate: Knowledge graph embedding by relational rotation in complex space," in *Proc. Int. Conf. Learn. Represent.*, 2019.
[52] D. Szklarczyk, J. H. Morris, H. Cook, M. Kuhn, S. Wyder, M. Simonovic, A. Santos, N. T. Doncheva, A. Roth, P. Bork, L. J. Jensen, and C. von Mering, "The STRING database in 2017: Quality-controlled protein–protein association networks, made broadly accessible," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D362–D368, 2017, doi: 10.1093/nar/gkw937.
[53] G. Tartari and A. Hogan, "WiSP: Weighted shortest paths for RDF graphs," in *Proc. VOILA@ISWC*, 2018.
[54] T. Trouillon, J. Welbl, S. Riedel, É. Gaussier, and G. Bouchard, "Complex embeddings for simple link prediction," in *Proc. ICML*, 2016, pp. 2071–2080.
[55] O. Udrea, D. R. Recupero, and V. S. Subrahmanian, "Annotated RDF," in *Proc. 3rd Eur. Semantic Web Conf.* (Lecture Notes in Computer Science), vol. 4011. Budva, Montenegro: Springer, Jun. 2006, pp. 487–501.
[56] O. Udrea, D. R. Recupero, and V. S. Subrahmanian, "Annotated RDF," *ACM Trans. Comput. Log.*, vol. 11, no. 2, p. 10, 2010.
[57] S. Vahdati, N. Arndt, S. Auer, and C. Lange, "OpenResearch: Collaborative management of scholarly communication metadata," in *Knowledge Engineering and Knowledge Management* (Lecture Notes in Computer Science), vol. 10024, E. Blomqvist, P. Ciancarini, F. Poggi, and F. Vitali, Eds. Cham, Switzerland: Springer, 2016, doi: 10.1007/978-3-319-49004-5_50.
[58] S. Vahdati, G. Palma, R. J. Nath, C. Lange, S. Auer, and M.-E. Vidal, "Unveiling scholarly communities over knowledge graphs," in *Digital Libraries for Open Knowledge* (Lecture Notes in Computer Science), vol. 11057, E. Méndez, F. Crestani, C. Ribeiro, G. David, and J. Lopes, Eds. Cham, Switzerland: Springer, 2018, doi: 10.1007/978-3-030-00066-0_9.
[59] K. Wang, Z. Shen, C. Huang, C.-H. Wu, Y. Dong, and A. Kanakia, "Microsoft academic graph: When experts are not enough," *Quant. Sci. Stud.*, vol. 1, no. 1, pp. 396–413, Feb. 2020.
[60] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM Comput. Surv.*, vol. 53, no. 3, pp. 1–34, 2020.
[61] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 1, pp. 4–24, Jan. 2021.
[62] C. Xu, M. Nayyeri, Y.-Y. Chen, and J. Lehmann, "Knowledge graph embeddings in geometric algebras," in *Proc. 28th Int. Conf. Comput. Linguistics*, 2020, pp. 530–544.
[63] B. Yang, W.-t. Yih, X. He, J. Gao, and L. Deng, "Embedding entities and relations for learning and inference in knowledge bases," 2014, *arXiv:1412.6575*. [Online]. Available: http://arxiv.org/abs/1412.6575
[64] L. Yao, Y. Zhang, B. Wei, Z. Jin, R. Zhang, Y. Zhang, and Q. Chen, "Incorporating knowledge graph embeddings into topic modeling," in *Proc. 31st AAAI Conf. Artif. Intell. (AAAI)*. AAAI Press, 2017, pp. 3119–3126.
[65] S. Zhang, Y. Tay, L. Yao, and Q. Liu, "Quaternion knowledge graph embeddings," 2019, *arXiv:1904.10281*. [Online]. Available: http://arxiv.org/abs/1904.10281
[66] J. Zhou, G. Cui, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, "Graph neural networks: A review of methods and applications," 2018, *arXiv:1812.08434*. [Online]. Available: https://arxiv.org/abs/1812.08434
[67] X. Zhou, Q. Zhu, P. Liu, and L. Guo, "Learning knowledge embeddings by combining limit-based scoring loss," in *Proc. ACM Conf. Inf. Knowl. Manage.*, Nov. 2017, pp. 1009–1018.

MOJTABA NAYYERI received the B.S. and M.S. degrees in computer engineering from Ferdowsi University of Mashhad, Mashhad, Iran, in 2011 and 2014, respectively. He is currently pursuing the Ph.D. degree with the Smart Data Analysis Group, University of Bonn, Germany. He has been a Research Assistant with the Nature-Inspired Machine Intelligence Research Group, Institute of Applied Informatics (InfAI), since 2020. His current research interests include machine learning, knowledge graphs, pattern recognition, and the semantic web.

GÖKCE MÜGE CIL received the B.Sc. degree in computer engineering from Bilkent University, Turkey, in 2017. She is currently pursuing the master's degree in computer science with the University of Bonn. She is working on her master's thesis with the Smart Data Analysis Group.



SAHAR VAHDATI received the M.Sc. and Ph.D. degrees in computer science from the University of Bonn. She has been a Senior Researcher and holds a postdoctoral position with Oxford University, U.K. She is currently leading the Nature-Inspired Machine Intelligence Research Group, Institute of Applied Informatics (InfAI), University Leipzig. Her research interests include using knowledge representation, analyzing knowledge graphs, and artificial intelligence (AI).

FRANCESCO OSBORNE is currently a Research Fellow with the Knowledge Media Institute, The Open University, U.K., where he leads the Scholarly Data Mining Team. His research interests include artificial intelligence, information extraction, knowledge graphs, the science of science, and the semantic web. He has authored more than 80 peer-reviewed publications in top journals and conferences in these fields. He collaborates with major publishers, universities, and companies in the space of innovation for producing a variety of innovative services for supporting researchers, editors, and research polities makers. He recently released the Computer Science Ontology that is currently the largest taxonomy of research areas in the field.

ANDREY KRAVCHENKO is currently a Researcher with the University of Oxford and with Skolkovo Institute of Science and Technology. His Ph.D. research was at the intersection of machine learning and unstructured data extraction. He also played a significant role in the DIADEM project, which produced state-of-the-art research in the field of large-scale fully automated web data extraction. His current research interests include theory and application of anomaly detection in big data using sequences and graphs, and in particular, the development of efficient machine learning algorithms based on the embedding of vectors. He works on exploring the broader connection between black-box machine learning models and knowledge-based systems, with a particular focus on knowledge graphs.

SIMONE ANGIONI received the B.S. and M.S. degrees in computer science from the University of Cagliari, Italy, where he is currently pursuing the Ph.D. degree. His research interests include the science of science, scientometrics, information extraction, the semantic web, and robotics. He is the main developer of the academia/industry dynamics (AIDA) knowledge graph, an innovative resource for studying the relationship between academia and industry.

ANGELO SALATINO received the Ph.D. degree in early detection of research trends. He is currently a Research Associate with the Intelligence Systems and Data Science (ISDS) Group, Knowledge Media Institute (KMi), The Open University. In particular, his project aimed at identifying the emergence of new research topics at their embryonic stage. His research interests include the semantic web, network science, and knowledge discovery technologies, with a focus on the structures and evolution of science.

DIEGO REFORGIATO RECUPERO received the Ph.D. degree in computer science from the University of Naples Federico II, Italy, in 2004. He has been an Associate Professor with the Department of Mathematics and Computer Science, University of Cagliari, Italy, since December 2016. From 2005 to 2008, he has been a Postdoctoral Researcher with the University of Maryland, College Park, USA. He is the author of more than 140 conference and journal papers in these research fields, with more than 1600 citations. His current research interests include sentiment analysis, semantic web, natural language processing, human–robot interaction, financial technology, and smart grid. He won different awards in his career (such as Marie Curie International Reintegration Grant, Marie Curie Innovative Training Network, Best Research Award from the University of Catania, Computer World Horizon Award, Telecom Working Capital, and Startup Weekend). He co-founded six companies within the ICT sector and is actively involved in European projects and research (with one of his companies, he won more than 40 FP7 and H2020 projects). In all of them, machine learning, deep learning, big data are key technologies employed to effectively solve several tasks.

ENRICO MOTTA received the Laurea degree in computer science from the University of Pisa, Italy, and the Ph.D. degree in artificial intelligence from The Open University. He is currently a Professor in knowledge technologies and the Former Director of the Knowledge Media Institute (KMi), The Open University, U.K., from 2000 to 2007. His research interests include the intersection of large-scale data integration and modeling, semantic and language technologies, intelligent systems, and human–computer interaction. Over the years, he has led KMi's contribution to numerous high-profile projects, receiving over £10.4M in external funding, since 2000, from a variety of institutional funding bodies and commercial organizations.

JENS LEHMANN received the Ph.D. degree *(summa cum laude)* from the University of Leipzig and the joint master's degree in computer science from the Technical University of Dresden and the University of Bristol. He is currently the Head of the Smart Data Analysis Research Group, a Full Professor with the University of Bonn, and a Lead Scientist with Fraunhofer IAIS. He authored more than 100 publications, which were cited more than 18000 times and have won 12 international awards. His research interests include semantic web technologies, question answering, machine learning, and knowledge graph analysis. He contributed to various open-source projects such as DL-Learner, SANSA, LinkedGeoData, and DBpedia.