Hierarchically Metric-Structured Knowledge Graph Embeddings

Anonymous authors
Paper under double-blind review

Abstract

In the vast landscape of big data, there is an important challenge in understanding data and structuring it in a suitable format. Knowledge graphs are considered a sophisticated solution to organize and infer data and knowledge, offering a structured framework that transcends disciplinary boundaries in medicine, culture, biology, social networks, music, and beyond. Despite their informativeness, these systems are typically incomplete and their intrinsic structure unknown, whereas existing methodologies for predicting missing facts and characterizing their structure face scalability and interpretability issues. Addressing this gap, we introduce a new latent feature model, leveraging the prominent RESCAL framework to account for degree heterogeneity, multiscale structure, and scalable inference using an approximation of the full likelihood of all triplets circumventing negative sampling inference strategies. This not only enhances computational efficiency but also provides deeper insights into the intrinsic multiscale structure of knowledge graphs, thereby advancing the interpretability of predictive models and paving the way for a more comprehensive understanding of complex information networks.

1 Introduction

Graphs, as a fundamental data structure, find application across diverse domains, including music, medicine, social networks, and more (Newman, 2003). The versatility of graphs is reflected in various structural manifestations such as unipartite graphs, bipartite graphs, and higher-order graphs such as knowledge graphs. A knowledge graph serves as a structured repository of information, encapsulating entities, relations, and semantic descriptions. Typically characterized by triples, these graphs represent factual relationships between two entities, thereby offering a comprehensive understanding of interconnected data. Some of the well-known curated knowledge graphs are YAGO (Suchanek et al., 2007), DBpedia (Auer et al., 2007), NELL (Carlson et al., 2010), Freebase (Bollacker et al., 2008), and Google Knowledge Graph (Google, May 2012). Despite the richness of information encapsulated within knowledge graphs, their inherent complexity surpasses conventional networks. This complexity necessitates the development of algorithms capable of mitigating modeling intricacies, especially when considering large-scale applications.

An often encountered challenge with knowledge graphs is their tendency to be incomplete and sparse, often lacking accurate factual information. A primary focus lies in the task of link prediction or knowledge graph completion, crucial for inferring missing connections within graphs and thereby completing the knowledge representation (Ali et al., 2021). The many research efforts developing and evaluating link prediction algorithms, particularly in real-world scenarios marked by sparsity, have underscored the potential of such methods to enhance the overall quality and utility of knowledge graphs significantly.

The world of knowledge graph completion involves different methods (Wang et al., 2021), which can be broadly classified into three categories. The first category utilizes the inherent *structure* of the knowledge graph itself, the second relies on *contextual* (side) information, and the third employs a *hybrid* approach that combines both. Within the *structure*-based methods, a further distinction can be made based on utilizing latent properties learned from the graph (not directly observed in the data), graph characteristics, or a combination of both (Nickel et al., 2015).

Structural methodologies often employ tensor factorization techniques, specifically those rooted in latent feature models. This includes notable methods such as the Bayesian Clustered Tensor Factorization (Sutskever et al., 2009), TOPHITS (Kolda et al., 2005), PITF (Rendle & Schmidt-Thieme, 2010), RESCAL (Nickel et al., 2011; Krompaß et al., 2013), TuckerER (Balažević et al., 2019) to mention but a few. Matrix factorization also plays a foundational role in this domain, contributing to a diverse range of methodologies (Jiang et al., 2012; Riedel et al., 2013; Huang et al., 2014). Unlike tensor factorization approaches, matrix factorization techniques rearrange the representation of a knowledge graph as a matrix such that for instance the rows could symbolize the subject-object pairs of the triples, while the columns signify the relations within the graph. Additionally, certain models adopt a neural network-based approach, employing multilayer perceptions (MLPs) such as E-MLP and ER-MLP, along with the Neural Tensor Network (NTN) (Socher et al., 2013; Dong et al., 2014). In contrast to tensor and matrix-based factorization techniques, neural network-based methods leverage hidden layers to encapsulate the intricate non-linear relationships between entities and relations.

Another category of models introduces latent distance considerations, determining the likelihood of relationships by analyzing distances between latent representations of entities. Examples within this category include Structured Embeddings (SE) (Bordes et al., 2011), TransE (Bordes et al., 2013), and RotateE (Sun et al., 2019). Furthermore, some models leverage graph features, such as ALEPH (Muggleton, 1995) and Path Ranking Algorithm (Lao & Cohen, 2010), whereas models have also been developed that combine the strengths of latent and graph-based approaches, as seen in ARE (Nickel et al., 2014) and stacking (Wolpert, 1992). These distance- and feature-based paradigms, and their evolution in knowledge graph embedding and completion, are comprehensively reviewed in the recent survey by Cao et al. (2024).

Navigating to contextual methodologies, the advent of Graph Neural Networks (GNNs) has significantly influenced knowledge graph completion. Notable models, such as the Relational Graph Convolutional Network (R-GCN) (Schlichtkrull et al., 2018), GraIL (Teru et al., 2020), and Decagon (Zitnik et al., 2018), have demonstrated the efficacy of this paradigm. Language models such as SimKGC (Wang et al., 2022) and KG-BERT (Yao et al., 2019) have also emerged, showcasing their effectiveness in capturing relevant contextual information. The integration of pretrained language models has become increasingly relevant, where multilingual and text-augmented frameworks (Song et al., 2024; Jiang et al., 2023) enhance the contextual grounding of entity-relation representations.

One of the important issues with knowledge graph completion is that knowledge graphs usually have only positive samples (i.e., triplets). Conventional methods often adopt a closed-world assumption, treating all non-positive triplets as negative, yet this may lead to scalability issues due to the potentially large number of false facts. Alternatively, some approaches use known constraints on the structure of a knowledge graph to generate more informative negative samples. However, how to devise the negative sampling procedure is non-trivial and can influence the learned representations, see also Nickel et al. (2015); Mishra et al. (2023) for a discussion. In our approach, we mitigate scalability concerns by integrating a hierarchical multi-scale approximation into the training process (Nakis et al., 2023) within the closed-world assumption framework. This strategy effectively provides an accurate estimation of the full likelihood of our model and thereby substantially reduces the computational overhead associated with evaluating the full likelihood for accurate learning while circumventing the need for negative sampling strategies.

Knowledge graphs are structured data that can be described further with semantic information (Stamou & Chortaras, 2017). Previous studies have utilized techniques such as latent class modeling approaches partitioning entities into concepts (Kemp et al., 2006) as well as inferring concept hierarchies through hierarchical clustering approaches to organize entities structured using the learned hierarchical organization (Roy et al., 2006; Nickel & Tresp, 2011; Pietrasik & Reformat, 2020), thereby offering a richer and more descriptive arrangement of data rooted in taxonomies. Our approach inherently provides a similar structure because of the hierarchical multi-scale approximation used in the training process. Thus, we naturally obtain a taxonomy of the entities within a knowledge graph as part of the inference.

In our investigation, we prioritize structural models due to their versatility in application to various knowledge graphs, eliminating the necessity for contextual information. To learn latent representations of relational data, we focus on RESCAL (Nickel et al., 2011; Krompaß et al., 2013), a method that leverages a

tensor factorization model designed to consider the inherent structure of relational data and has recently been confirmed as a current state-of-the-art structural model (Teach et al., 2020; Kong et al., 2019). We demonstrate how the RESCAL model can be reparameterized, transforming it into a distance model with random effects providing a seamless integration of hierarchical clustering into the training process while at the same time enabling to explicitly account for relation-specific degree heterogeneity. In this manner, we provide an accurate full-likelihood approximation, leveraging Nakis et al. (2023), importantly circumventing issues of negative sampling in conventional knowledge graph learning. This novel approach thereby allows us to capture the structural intricacies of a knowledge graph effectively. By defining latent representations for relation-specific nodes within the graph, we unveil compelling hierarchical clustering patterns, thereby enhancing our understanding of the underlying data in terms of the learned hierarchical (i.e., taxonomic) representation.
Specifically, our contributions are:

- We demonstrate how RESCAL can be reformulated as a latent distance model with random effects.
- We thereby address the challenge of negative sampling and provide a scalable accurate hierarchical full likelihood approximation exploring the metric properties of our reformulation.
- On a variety of benchmark knowledge graphs, we show how our proposed model provides accurate link prediction while at the same time admitting direct inference of concept hierarchies.

The paper is organized as follows. In Section 2, we present the RESCAL procedure and introduce the proposed Hierarchically Metric Structured Knowledge Graph Embeddings (HMSKGE) method. Section 3 details and discuss the results of our models on two tasks: link prediction and taxonomy induction. Finally, Section 4 concludes the paper by summarizing the findings, addressing limitations, and exploring the broader impact of the work.

2 Methodology

We employed the Resource Description Framework (RDF) format from the semantic web to depict knowledge graphs, employing triples composed of (subject, predicate, and object) to denote relationships (Decker et al., 2000). Within this framework, the predicates denote connections between entities.

We conceptualized a knowledge graph as a tensor, denoted as $\mathcal{X} \in \{0,1\}^{N_e \times N_e \times N_r}$, where N_e represents the number of entities, and N_r signifies the number of relations involved.. Every element, denoted as $\mathcal{X}_{ijk}$, within the tensor, holds a value of one to indicate the existence of a relationship. This signifies that there is a connection between the $i-th$ and the $j-th$ entity regarding the $k-th$ predicate. Conversely, the entry is set to zero if a relationship does not exist or is unknown.

2.1 RESCAL

To learn latent representations of relational data, we focus on RESCAL (Nickel et al., 2011; Krompaß et al., 2013) - a current state-of-the-art structural model (Teach et al., 2020; Kong et al., 2019). According to the RESCAL model, each relational slice $\mathcal{X}_k$ of the tensor $\mathcal{X}$ is approximated by:

$$\mathcal{X}_k \approx \mathcal{E}\mathcal{R}_k\mathcal{E}^\top, for\ k = 1, 2, \ldots, N_r. \tag{1}$$

As a result, $\mathcal{E}$ represents a $N_e \times D$ matrix containing the latent component representation of entities within the domain, and $\mathcal{R}_k$ is an asymmetric $D \times D$ matrix that characterizes the interactions of latent components associated with the $k-th$ predicate corresponding to a slice of the tensor $\mathcal{R} \in \mathbb{R}^{D \times D \times N_r}$.

In the original work on RESCAL model, (Nickel et al., 2011) estimation was based on least squares minimization. However, given the binary nature of tensor $\mathcal{X}$, we estimated Equation 1 by assuming that $\mathcal{X}_{ijk}$ follows a Bernoulli distribution which was demonstrated in (Nickel & Tresp, 2013) to enhance performance.

This approach aligns with common practices in modeling network data, as highlighted in previous works (Hoff et al., 2002; Handcock et al., 2007; Krivitsky et al., 2009). Consequently, we define the BERNOULLI model and optimize $\mathcal{E}$ and $\mathcal{R}$ by employing stochastic gradient descent on the negative log-likelihood given by

$$\mathcal{L}(\mathcal{X}|\mathcal{E},\mathcal{R}) := \sum_{i,j,k} x_{ijk}\eta_{ijk} - \log(1+\exp(\eta_{ijk})), \quad \eta_{ijk} := \mathbf{e}_i\mathcal{R}_k\mathbf{e}_j^\top. \tag{2}$$

where $\mathbf{e}_i \in \mathbb{R}^D$ represent the embedding of node i.

2.2 The Hierarchically Metric Structured Knowledge Graph Embeddings

To enhance the model's versatility, we integrate random effects into the likelihood, thereby formulating the BERNOULLI-RE model. This extension is seamlessly achieved by adding β_{ik} and γ_{jk}, in the definition of η_{ijk} and enables direct modeling of degree heterogeneity across the predicates.

Prior studies (Wind & Mørup, 2012; Nakis et al., 2023) have shown that incorporating a Poisson likelihood given by

$$\mathcal{L}(\mathcal{X}|\mathcal{E},\mathcal{R},\beta,\gamma) := \sum_{i,j,k} x_{ijk}\log\lambda_{ijk} - \lambda_{ijk}, \quad \lambda_{ijk} := \exp\left(\mathbf{e}_i\mathcal{R}_k\mathbf{e}_j^\top + \beta_{ik} + \gamma_{jk}\right), \tag{3}$$

does not diminish a model's predictive accuracy and at the same time can be easily applied to weighted knowledge graphs. Additionally, Poisson models possess beneficial decoupling properties among predictor variables in the likelihood function (Karrer & Newman, 2011; Herlau et al., 2014; Nakis et al., 2023), which we utilize in the following to accurately approximate the full likelihood.

Our current modeling approach primarily utilizes pairwise latent interactions, where the likelihood of a triplet's realization is determined by evaluating $\mathbf{e}_i\mathcal{R}_k\mathbf{e}_j^\top$. Transitioning from the RESCAL model to a Latent Distance Model (LDM) (Hoff et al., 2002) is seamless when the model includes random effects and can be accomplished through reparameterization. In this modified model, the score assigned to each triplet is contingent on the distances between latent representations of entities and relations. Specifically, we defined the LDM model by setting λ_{ijk} of Equation 3 as:

$$\lambda_{ijk} := \exp(\hat{\beta}_{ik} + \hat{\gamma}_{jk} - \frac{1}{2}\|\mathbf{e}_i\mathcal{V}_k - \mathbf{e}_j\mathcal{Q}_k\|_2^2) \tag{4}$$

where $\mathcal{V}_k$ and $\mathcal{Q}_k$ correspond to the k^{th} slices of the tensors $\mathcal{V} \in \mathbb{R}^{D\times D\times N_r}$ and $\mathcal{Q} \in \mathbb{R}^{D\times D\times N_r}$ accounting for the relation-specific information such that $\mathcal{R}_k = \mathcal{V}_k\mathcal{Q}_k^\top$. Noting that

$$\|\mathbf{e}_i\mathcal{V}_k - \mathbf{e}_j\mathcal{Q}_k\|_2^2 = (\mathbf{e_i}\mathcal{V}_k - \mathbf{e}_j\mathcal{Q}_k)(\mathbf{e}_i\mathcal{V}_k - \mathbf{e}_j\mathcal{Q}_k)^\top = \mathbf{e}_i\mathcal{V}_k\mathcal{V}_k^\top\mathbf{e}_i^\top + \mathbf{e}_j\mathcal{Q}_k\mathcal{Q}_k^\top\mathbf{e}_j^\top - 2\mathbf{e}_i\mathcal{V}_k\mathcal{Q}_k^\top\mathbf{e}_j\top, \tag{5}$$

we obtain Equation 4 by defining $\beta_{ik} = \hat{\beta_{ik}} - \frac{1}{2}\mathbf{e}_i\mathcal{V}_k\mathcal{V}_k^\top\mathbf{e}_i^\top$ and $\gamma_{jk} = \hat{\gamma_{jk}} - \frac{1}{2}\mathbf{e}_j\mathcal{Q}_k\mathcal{Q}_k^\top\mathbf{e}_j^\top$.

Importantly, this reparameterization to a conventional LDM model (Equation 4) with random effects (β_{ik} and γ_{jk}) enables us to explore the hierarchical block approximation recently proposed in the context of the LDM model in Nakis et al. (2023). Our goal is to streamline the overall likelihood computation while structuring our entities into coherent hierarchies or categories. This is accomplished by constructing a block-like hierarchical arrangement using a clustering approach applied to latent variables in Euclidean space. We organized the embedded clusters into a hierarchy using a tree structure, defining a cluster dendrogram forming the Hierarchically Metric Structured Knowledge Graph Embeddings (HMSKGE) model. The HMSKGE is defined by the following loss function:

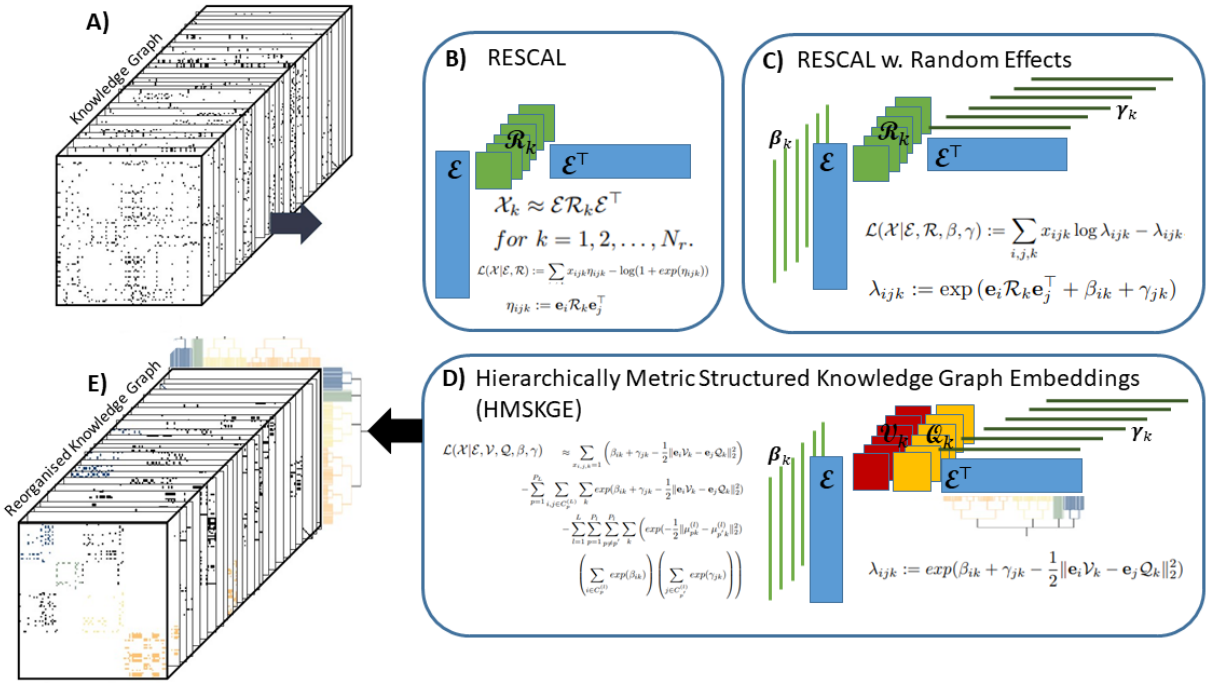


Figure 1: Illustration of the HMSKGE applied to the *Kinships* knowledge graph. **A)** The *Kinships* knowledge graph. **B)** The RESCAL model decomposes the knowledge graph into latent factors $\mathcal{E}$ as well as relation-specific interactions $\mathcal{R}_k$ originally solved using the least-squares loss but advanced to Bernoulli (cross-entropy) loss here illustrated with improved results. **C)** To explicitly account for degree heterogeneity we include random effects and explore that random effects efficiently decouple in the inference by utilizing the Poisson loss that can also account for integer-weighted graphs. **D)** Exploring the RESCAL formulation with random effects we show that the model can be reformulated as a distance model admitting efficient and accurate full likelihood approximation not relying on negative sampling procedures during inference. **E)** Visualization results when reorganizing the nodes according to the inferred hierarchical structure learned by organizing the knowledge graph in coherent groups at multiple scales of the inferred hierarchy.

$$
\begin{aligned}
\mathcal{L}(\mathcal{X}|\mathcal{E},\mathcal{V},\mathcal{Q},\beta,\gamma) &:= \sum_{x_{i,j,k}=1}\left(\beta_{ik}+\gamma_{jk}-\frac{1}{2}\|\mathbf{e}_i\mathcal{V}_k-\mathbf{e}_j\mathcal{Q}_k\|_2^2\right) - \sum_{i,j,k}\exp(\beta_{ik}+\gamma_{jk}-\frac{1}{2}\|\mathbf{e}_i\mathcal{V}_k-\mathbf{e}_j\mathcal{Q}_k\|_2^2) \\
&\approx \sum_{x_{i,j,k}=1}\left(\beta_{ik}+\gamma_{jk}-\frac{1}{2}\|\mathbf{e}_i\mathcal{V}_k-\mathbf{e}_j\mathcal{Q}_k\|_2^2\right) - \sum_{p=1}^{P_L}\sum_{i,j\in C_p^{(L)}}\sum_k \exp(\beta_{ik}+\gamma_{jk}-\frac{1}{2}\|\mathbf{e}_i\mathcal{V}_k-\mathbf{e}_j\mathcal{Q}_k\|_2^2) \\
&- \sum_{l=1}^{L}\sum_{p=1}^{P_l}\sum_{p'\neq p}^{P_l}\sum_k \left(\exp(-\frac{1}{2}\|\mu_{pk}^{(l)}-\mu_{p'k}^{(l)}\|_2^2)\left(\sum_{i\in C_p^{(l)}}\exp(\beta_{ik})\right)\left(\sum_{j\in C_{p'}^{(l)}}\exp(\gamma_{jk})\right)\right)
\end{aligned}
\tag{6}
$$

where $l \in \{1, \ldots, L\}$ denotes the lth dendrogram level, $p \in \{1, \ldots, P_l\}$ indexes the cluster $C_p^{(l)}$ at that level, and $\mu_{pk}^{(l)}$ is the corresponding centroid for the kth relation. An example of an application of our model is given in Figure 1.

2.2.1 Dendrogram construction

The process of dendrogram construction initiates by identifying clusters within its initial layer, typically with a count approximately equal to the $\ln N_e$. This initial clustering is facilitated through the utilization of the standard K-MEANS algorithm applied to all entities. Herein, the latent representation of each entity is defined as the concatenation of $\mathbf{e}_i\mathcal{V}_k$ and $\mathbf{e}_i\mathcal{Q}_k$ for all relations $k \in 1, \ldots, K$, i.e. $\mathbf{z}_i = [\mathbf{e}_i\mathcal{V}_1\ \mathbf{e}_i\mathcal{Q}_1, \ldots, \mathbf{e}_i\mathcal{V}_K\ \mathbf{e}_K\mathcal{Q}_K]$ allowing for a comprehensive consideration of the position of all entity relations across the predicates during the clustering procedure performed on $\mathbf{Z} \in \mathbb{R}^{N_e \times N_r \cdot 2 \cdot D}$. Consequently, centroids per relation and cluster can be derived and represented as $\mu_p \in \mathbb{R}^{N_r \cdot 2 \cdot D}$. Subsequently, every cluster is further divided through the application of the K-MEANS algorithm, resulting in the creation of two smaller clusters, effectively doubling the total count to 2 clusters. This subdivision process iterates until each cluster reaches a minimum threshold of $\ln(N_e)$ entities.

Crucially, the model formulation given in Equation 4 suffers from redundancies in the scales of $\mathbf{V}_k$ and $\mathbf{Q}_k$ such that the one can become large countered by the other becoming low in values and vice versa with the random effects also absorbing these changes in scale. When performing clustering in the space $\mathbf{z}_i = [\mathbf{e}_i\mathcal{V}_1\ \mathbf{e}_i\mathcal{Q}_1, \ldots, \mathbf{e}_i\mathcal{V}_K\ \mathbf{e}_K\mathcal{Q}_K]$ this in turns results in low variance components for some parts of the concatenation making the clustering pay little attention to these parts when performing the segmentation. To remedy this, we ensure by reparameterising the model that $\mathbf{V}_k$ and $\mathbf{Q}_k$ have same magnitude using the following transformations of the parameters that leaves Equation 4 invariant:

$$\tilde{\mathcal{V}}_k(d) = \frac{\sqrt{\|\mathcal{V}_k(d)\|_F \|\mathcal{Q}_k(d)\|_F}}{\|\mathcal{V}_k(d)\|_F}\mathcal{V}_k(d), \qquad \tilde{\mathcal{Q}}_k(d) = \frac{\sqrt{\|\mathcal{V}_k(d)\|_F \|\mathcal{Q}_k(d)\|_F}}{\|\mathcal{Q}_k(d)\|_F}\mathcal{Q}_k(d), \tag{6}$$

$$\tilde{\beta}_{ik} = \beta_{ik} - \frac{1}{2}(\mathbf{I} - \frac{\|\mathcal{Q}_k\|_F}{\|\mathcal{V}_k\|_F})\mathbf{e}_i\mathcal{V}_k\mathcal{V}_k^\top\mathbf{e}_i^\top, \qquad \tilde{\gamma}_{jk} = \gamma_{jk} - \frac{1}{2}(\mathbf{I} - \frac{\|\mathcal{V}_k\|_F}{\|\mathcal{Q}_k\|_F})\mathbf{e}_j\mathcal{Q}_k\mathcal{Q}_k^\top\mathbf{e}_j^\top. \tag{7}$$

Using these normalizations it can be trivially shown that

$$\exp(\beta_{ik} + \gamma_{jk} - \frac{1}{2}\|\mathbf{e}_i\mathcal{V}_k - \mathbf{e}_j\mathcal{Q}_k\|_2^2) = \exp(\tilde{\beta}_{ik} + \tilde{\gamma}_{jk} - \frac{1}{2}\|\mathbf{e}_i\tilde{\mathcal{V}}_k - \mathbf{e}_j\tilde{\mathcal{Q}}_k\|_2^2). \tag{8}$$

This re-normalization is applied at every iteration of the HMSKGE inference procedure.

2.2.2 Complexity

The computational complexity associated with estimating Equation 6 becomes linearithmic in the number of entities N_e when terminating at clusters of size $\ln(N_e)$, reducing the overall cost of evaluating the full likelihood from $\mathcal{O}(N_e^2 N_r)$ to $\mathcal{O}(N_e \log(N_e) N_r)$. Compared with classical tensor factorization approaches such as RESCAL_ALS (complexity $\mathcal{O}(N_r N_e D^2 + N_r D^3)$) and probabilistic formulations like POISSON, BERNOULLI, and LDM (complexity $\mathcal{O}((|\mathcal{E}| + C)D^2)$, where C denotes the number of negative samples), the proposed HMSKGE achieves a substantially lower complexity of $\mathcal{O}(N_e \log(N_e) N_r D^2)$. This hierarchical approximation eliminates the reliance on negative sampling, which, while common in existing methods, introduces stochastic variability and may neglect important entity interactions. By organizing entities according to their latent-space proximity, HMSKGE concentrates computational effort on local neighborhoods where accurate predictions are most critical, while concurrently producing an interpretable hierarchical taxonomy that reflects entity similarity. This combination of efficiency, scalability, and interpretability makes HMSKGE particularly well-suited for large-scale knowledge graph embedding and inference.

2.2.3 Embedding Proximity

The likelihood of a triplet in a knowledge graph can be related to the distance between the embeddings of the subject and object entities. Intuitively, entities that are closer in the embedding space are more likely to be connected. Lemma 1 formalizes this intuition by relating embedding proximity to the probability of a triplet.

Lemma 1. *Let $\{\mathcal{X}_{ijk}\}_{(i,j,k)\in N_e\times N_e\times N_r}$ be the set of independent random variables indicating the presence of edges and following a Poisson distribution. Then, each triple (i,j,k) satisfies*

$$-\log\left(\epsilon p_{ijk}\right)+\beta_{ik}+\gamma_{jk}\geq\frac{1}{2}\|\mathbf{e}_i\mathcal{V}_k-\mathbf{e}_j\mathcal{Q}_k\|_2^2 \tag{9}$$

where $p_{ijk}=P(\mathcal{X}_{ijk}>\epsilon)$ and ϵ denotes the weight of the interaction.

Proof. Since we suppose that each triple is an independent and identically distributed variable, by Markov's inequality, we can write that

$$P(\mathcal{X}_{ijk}>\epsilon)\leq\frac{\mathbb{E}[\mathcal{X}_{ijk}]}{\epsilon}=\frac{\lambda_{ijk}}{\epsilon}=\frac{\exp\left(\beta_{ik}+\gamma_{jk}-\frac{1}{2}\|\mathbf{e}_i\mathcal{V}_k-\mathbf{e}_j\mathcal{Q}_k\|_2^2\right)}{\epsilon} \tag{10}$$

which implies that

$$-\log\left(\epsilon p_{ijk}\right)+\beta_{ik}+\gamma_{jk}\geq\frac{1}{2}\|\mathbf{e}_i\mathcal{V}_k-\mathbf{e}_j\mathcal{Q}_k\|_2^2 \tag{11}$$

□

3 Results & Discussion

We conducted a series of experiments to evaluate the expressiveness of the embeddings generated by the HMSKGE model. These experiments were carried out on a Tesla A100 PCIE 80 GB GPU, and focused on link prediction as well as taxonomy induction. We used a diverse range of datasets of small and large sizes and compared our results against multiple baseline models to ensure a thorough evaluation.

The datasets employed in this study encompass *Kinships* (Quinlan, 1990), the *WN18RR* dataset (Dettmers et al., 2018), the *FB15k-237* dataset (Toutanova & Chen, 2015), and the *YAGO3-10* dataset (Dettmers et al., 2018). The *Kinships* knowledge graph specifically delineates kinship relations within the Alwayarra tribe, as documented by (Kemp et al., 2006). The *WN18RR* dataset is derived from the *WN18* subset of WordNet, while the *FB15k-237* dataset is a modified version of the *FB15k* dataset. The *FB15k-237* dataset is specifically designed based on Freebase entity pairs, eliminating inverse relations to address concerns related to the impact of test triplets generated by inverting training set triplets. The *YAGO3-10* is a subset of the *YAGO3* dataset that combines the information from Wikipedia articles in multiple languages with WordNet, GeoNames, and other data sources. The above datasets have been widely used as benchmarks for the link prediction task. Table 1 provides the statistics of all the datasets that we used.

Table 1: Statistics of the datasets considered in this work.

Dataset	**Triples**	**Entities**	**Relations**	**Mean-degree**	**Median-degree**
YAGO3-10	1,179,040	123,182	37	12.640	8
FB15K-237	310,116	14,541	237	34.194	21
WN18RR	93,003	40,943	11	3.727	3
Kinships	10,790	104	26	105	105

For the baseline model, we selected RESCAL (Nickel et al., 2011), implemented using alternating least squares optimization during training. To systematically evaluate the proposed HMSKGE, we contrasted its derivation steps against the RESCAL baseline using multiple formulations: the Bernoulli likelihood (Bernoulli), the Bernoulli likelihood with random effects (Bernoulli-RE), and the Poisson likelihood (Poisson). We further considered a Poisson-based reformulation as a distance model, which we denote LDM.

All baseline models were trained using batching and negative sampling (Dettmers et al., 2018). This setup enabled us to quantify the performance gap between the hierarchical approximation of the full likelihood

and conventional inference with negative sampling, using both the LDM and Poisson formulations. Implementation details and code are available in the GitHub repository[1].

3.1 Link Prediction

For the task of link prediction, also referred to as knowledge graph completion, we present results using two evaluation metrics: one rank-based metrics, as proposed by Ali et al. (2021); Hoyt et al. (2022), and one classification-based metric. The rank-based metric used is HITS@10, which evaluates the performance of ranking systems by measuring the proportion of true entities that appear within the top 10 ranks, with higher values indicating better performance. The classification-based metric, AUC-ROC, summarize the model's ability to discriminate between positive and negative instances across all classification thresholds. The standard train–test split provided with each benchmark dataset was used, consistent with prior work. For the ranking-based evaluation, the model was trained on the training set and evaluated on the test set by performing head and tail prediction, as is standard in knowledge graph completion. Specifically, for each positive triplet in the test set, we replaced the head (or tail) entity in turn with every possible entity in the graph and ranked all resulting candidate triplets according to their predicted scores. The true entity's rank was then recorded, and the HITS@10 metric was computed as the average proportion of cases where the correct entity appeared within the top 10 ranks, averaged over both head and tail prediction tasks. For the AUC-ROC evaluation, we randomly created negative test samples following a balanced sampling strategy: an equal number of negative triplets were generated by randomly replacing either the head or tail entity in the positive test triplets, ensuring that the resulting triplets did not appear in the training or test sets. The model then produced a score for each positive and negative triplet using the learned embeddings, and the AUC-ROC was computed based on these scores and their corresponding binary labels.

In Tables 2 and 3, we report the results for Rescal, HMSKGE, and the corresponding ablation models used as baselines. All models were trained with embedding dimensions of 2, 3 and 16, using a learning rate of 0.001 over 100,000 epochs. Our models consistently outperform the traditionally trained Rescal, which relies on alternating least squares optimization, confirming that gradient descent optimization based on Bernoulli likelihoods yields superior performance, consistent with prior findings (Nickel & Tresp, 2013). Incorporating random effects further improves performance, likely by better capturing nodes that act as hubs in the network and efficiently characterizing degree heterogeneity within each relational layer. The Bernoulli and Poisson variants exhibit similar behavior, in line with previous work (Wind & Mørup, 2012), and the Poisson and LDM models also show closely aligned performance patterns. The HMSKGE model performs strongly in the link prediction task, with only a slight deterioration in performance, probably due to the inherent approximation introduced by the hierarchical modeling. Notably, the HITS@10 and AUC-ROC metrics exhibit distinct behaviors. While AUC-ROC demonstrates a strong ability to distinguish between model performances—particularly in large-scale or imbalanced networks—HITS@10 can be informative for specific ranking tasks but may overlook subtle differences and is less robust when used in isolation. Therefore, combining complementary metrics such as AUC-ROC or HITS@k is recommended to ensure a more comprehensive and reliable evaluation of link prediction performance.

Based on the results in Tables 2 and 3, where HMSKGE showed limited performance in low-dimensional settings, we further evaluated RESCAL, HMSKGE, and our ablation baselines under varying large embedding sizes and training parameters (Tables 4 and 5). All models were trained with embedding dimensions of 20, and 30, a learning rate of 0.01, and for 20,000 epochs—except for WN18RR, which required 100,000 epochs and a reduced learning rate of 0.001. We used a smaller learning rate for WN18RR because its complex relational structure made training unstable at 0.01; lowering it to 0.001 ensured smoother convergence during extended training. The extended results confirm that while our approach underperforms in low-dimensional spaces, it achieves substantial gains as the embedding dimension increases, likely because higher-dimensional representations better capture hierarchical structures in the latent space.

In summary, the results in Tables 2, 3, 4 and 5 indicate that while HMSKGE is effective at capturing relational patterns in knowledge graphs, its performance is sensitive to the embedding dimensionality, with higher dimensions enabling the model to better capture hierarchical structures used for the hierarchical

[1]Not provided due to the double-blind review policy; provided as supplementary material.

Table 2: Comparison of HITS@10 and AUC-ROC scores for models on the *Kinships* and *WN18RR* datasets averaged over 3 runs. The standard error of the mean for all cases is approximately 0.001.

Model	Metric (D)	*Kinships* 2	*Kinships* 3	*Kinships* 16	*WN18RR* 2	*WN18RR* 3	*WN18RR* 16
RESCAL	HITS@10	0.100	0.721	**0.967**	0.013	0.024	0.279
	AUC-ROC	0.959	0.977	**0.974**	0.788	0.849	0.849
Bernoulli	HITS@10	0.500	0.701	0.955	0.029	0.039	0.404
	AUC-ROC	0.971	0.956	0.965	0.923	0.938	0.899
Bernoulli-RE	HITS@10	0.698	0.854	0.937	**0.043**	0.044	0.407
	AUC-ROC	0.958	0.945	0.955	0.940	**0.944**	0.891
Poisson	HITS@10	0.828	**0.889**	0.940	0.043	**0.046**	**0.429**
	AUC-ROC	0.958	0.947	0.955	**0.941**	**0.944**	0.885
LDM	HITS@10	**0.877**	0.833	0.946	0.038	0.042	0.386
	AUC-ROC	0.948	0.961	0.957	0.922	0.942	0.866
HMSKGE	HITS@10	0.203	0.196	0.673	0.017	0.004	0.162
	AUC-ROC	0.916	0.913	0.918	0.784	0.765	**0.920**

Table 3: Comparison of HITS@10 and AUC-ROC scores for models on the *FBK15K-237* and *YAGO3-10* datasets averaged over 3 runs. The standard error of the mean for all cases is approximately 0.001.

Model	Metric (D)	*FBK15K-237* 2	*FBK15K-237* 3	*FBK15K-237* 16	*YAGO3-10* 2	*YAGO3-10* 3	*YAGO3-10* 16
RESCAL	HITS@10	0.126	0.132	0.142	0.022	0.021	0.010
	AUC-ROC	0.932	0.954	0.967	0.908	0.910	0.915
Bernoulli	HITS@10	0.138	0.265	0.448	0.049	0.031	**0.068**
	AUC-ROC	0.993	0.996	**0.999**	0.980	0.901	**0.997**
Bernoulli-RE	HITS@10	0.133	0.348	0.448	**0.058**	0.058	0.067
	AUC-ROC	0.995	**0.997**	**0.999**	**0.986**	0.987	**0.997**
Poisson	HITS@10	**0.326**	**0.352**	**0.449**	**0.058**	0.058	0.064
	AUC-ROC	**0.996**	**0.997**	0.998	**0.986**	0.986	0.995
LDM	HITS@10	0.277	0.314	0.425	0.054	**0.059**	0.045
	AUC-ROC	0.986	**0.997**	0.996	0.981	**0.990**	0.965
HMSKGE	HITS@10	0.112	0.110	0.226	0.014	0.008	0.032
	AUC-ROC	0.843	0.899	0.996	0.852	0.899	0.938

approximation of the full likelihood. The experiments also show that gradient-based optimization of likelihoods provides a clear advantage over traditional alternating least squares methods, and that including random effects can further refine model performance. Moreover, rank-based metrics such as HITS@10 prove to be more discriminative for assessing link prediction than AUC-ROC. Overall, the findings highlight the trade-offs between model capacity, dimensionality, and metric choice in achieving accurate link prediction in diverse knowledge graphs.

3.2 Taxonomy Induction

As described in Pietrasik & Reformat (2020), information in knowledge graphs is typically structured using an ontology, which provides semantics to the knowledge graph's triplets through an axiomatic foundation

Table 4: Comparison of HITS@10 and AUC-ROC scores for *Kinships* and *WN18RR* datasets averaged over 3 runs. The standard error of the mean for all cases is approximately 0.005.

	Metric	*Kinships*			*WN18RR*		
Model	(D)	10	20	30	10	20	30
RESCAL	HITS@10	0.924	**0.975**	**0.970**	0.029	0.049	0.060
	AUC-ROC	0.959	**0.977**	**0.974**	0.616	0.731	0.752
Bernoulli	HITS@10	**0.964**	0.930	0.870	0.362	**0.415**	**0.424**
	AUC-ROC	**0.971**	0.956	0.946	0.955	0.895	0.872
Bernoulli-RE	HITS@10	0.939	0.900	0.855	0.323	0.395	0.416
	AUC-ROC	0.958	0.945	0.935	**0.963**	**0.957**	**0.952**
Poisson	HITS@10	0.941	0.921	0.862	0.325	0.397	0.418
	AUC-ROC	0.958	0.947	0.932	0.932	0.901	0.893
LDM	HITS@10	0.948	0.950	0.943	**0.365**	0.346	0.356
	AUC-ROC	0.948	0.961	0.958	0.847	0.911	0.843
HMSKGE	HITS@10	0.667	0.678	0.647	0.028	0.111	0.112
	AUC-ROC	0.916	0.913	0.910	0.865	0.917	0.917

Table 5: Comparison of HITS@10 and AUC-ROC scores for *FBK15K-237* and *YAGO3-10* datasets averaged over 3 runs. The standard error of the mean for all cases is approximately 0.005.

	Metric	*FBK15K-237*			*YAGO3-10*		
Model	(D)	10	20	30	10	20	30
RESCAL	HITS@10	0.208	0.335	0.366	**0.105**	0.122	0.143
	AUC-ROC	0.841	0.898	0.893	0.942	0.946	0.953
Bernoulli	HITS@10	0.438	0.438	0.407	0.067	0.065	0.072
	AUC-ROC	**0.999**	0.996	0.990	**0.997**	0.993	0.993
Bernoulli-RE	HITS@10	**0.442**	**0.439**	0.407	0.067	0.075	0.092
	AUC-ROC	0.942	0.877	0.859	**0.997**	0.996	0.995
Poisson	HITS@10	0.439	0.445	**0.439**	0.090	**0.220**	**0.248**
	AUC-ROC	**0.999**	**0.998**	**0.996**	**0.997**	**0.998**	**0.998**
LDM	HITS@10	0.407	0.430	0.434	0.057	0.070	0.114
	AUC-ROC	0.996	0.991	0.986	0.996	0.996	0.996
HMSKGE	HITS@10	0.227	0.254	0.276	0.032	0.034	0.042
	AUC-ROC	0.995	0.995	0.943	0.938	0.929	0.935

that defines the associations between entities and relations. A crucial component of most ontologies is the class taxonomy, organized through class subsumption axioms that represent is-a relationships between classes. This taxonomy often resembles a rooted tree, with a root class from which all other classes logically descend. The challenge of class taxonomy induction from knowledge graphs involves generating subsumption axioms from triplets to construct the class taxonomy.

To evaluate the clustering performance of the model, we applied the HMSKGE model with an embedding dimension of 64 to the *Kinships* knowledge graph. The resulting inferred taxonomy is shown in Figure 2. The *Kinships* graph represents a tribe in Australia, organized into four distinct communities. It contains metadata for each individual and defines 26 different types of relationships among tribe members. The

inferred taxonomy exhibits a clear hierarchical organization: the first level clusters individuals by tribe affiliation, the second by sex, and the third by age. This example demonstrates that the hierarchical structure learned by the HMSKGE model effectively captures and organizes the underlying relational patterns in the knowledge graph.

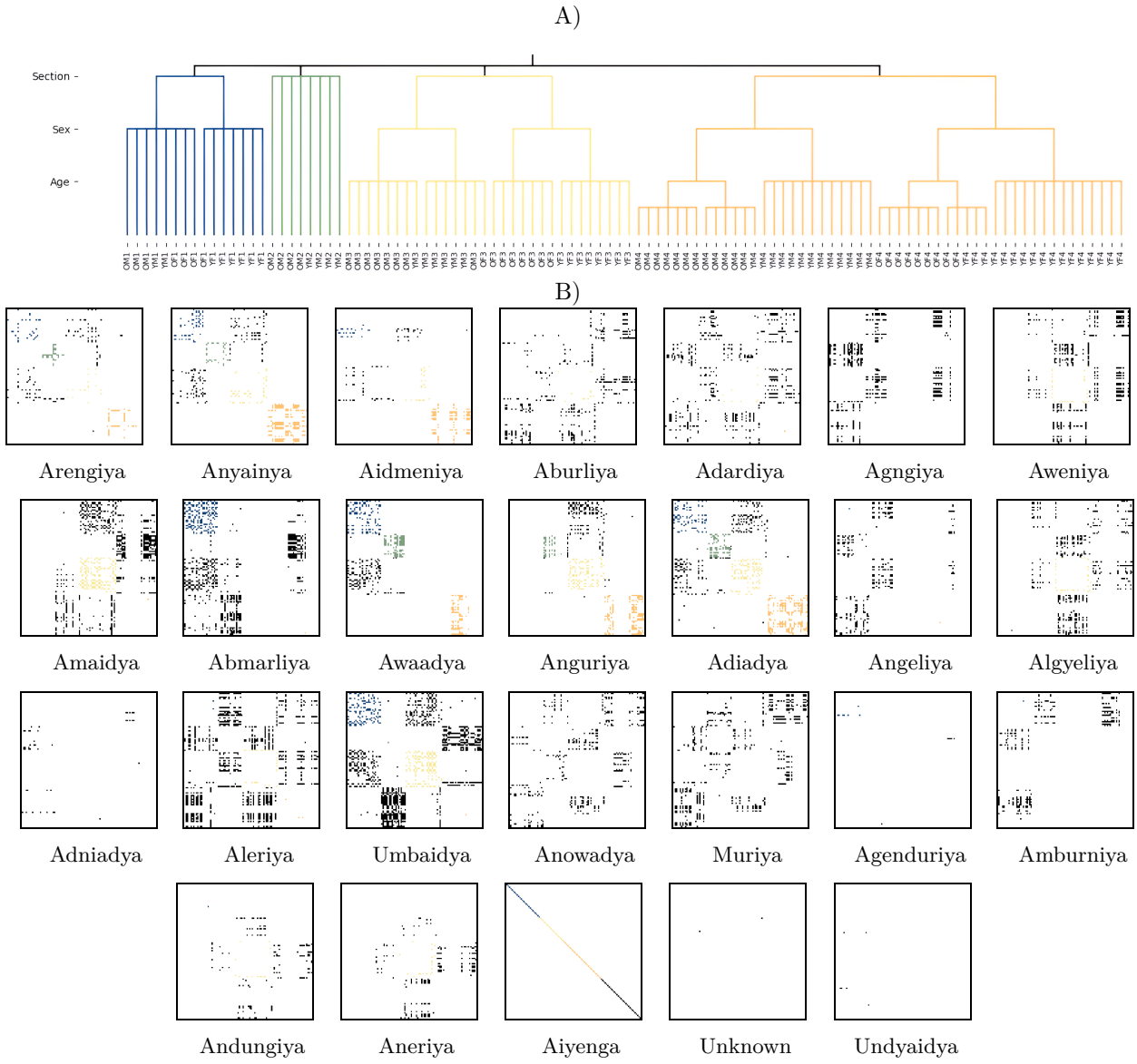


Figure 2: **A)** The dendrogram illustrates the *Kinships* dataset, where 'O' denotes old age, 'Y' denotes young age, 'M' denotes male, 'F' denotes female, and numbers represent different families within the tribe. **B)** The adjacency matrices per relation in the *Kinships* knowledge graph, reordered according to the inferred hierarchy given at the top.

We additionally visualize the smallest *WN18RR* of the larger knowledge graph providing a complementary perspective on the clustering behavior of the HMSKGE model for the visualization configured with an embedding dimension of 16. As shown in Figure 3, the adjacency matrices corresponding to different relation types exhibit clear block structures, indicating that nodes sharing similar semantic or relational roles are grouped together. Some faint lines can also be observed across several matrices; these appear to result from the random effects inherent to our model, which in turn enable it to capture hub nodes accurately.

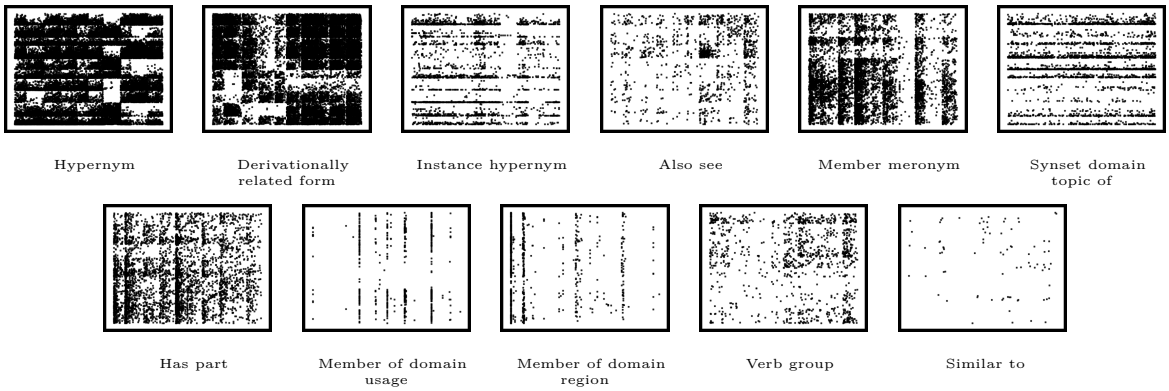


Figure 3: Adjacency matrices for each relation in the *WN18RR* knowledge graph, reordered according to clusters inferred by HMSKGE. Each matrix shows the connectivity pattern for a specific relation, with entities arranged to highlight relationally coherent clusters. Clear block structures indicate that entities sharing similar relational roles are grouped together.

Overall, these results confirm that HMSKGE effectively captures hierarchical and relational structures in knowledge graphs. By leveraging both entity interactions and relation-specific patterns, the model produces interpretable clusters and taxonomies that reflect meaningful latent groupings. Visualizations of adjacency matrices and dendrograms further illustrate how entities are organized according to their relational roles, providing clear insights into the structure of the graphs. While these results are dataset-specific, they highlight the HMSKGE model's ability to uncover interpretable hierarchies and clusters in knowledge graphs.

4 Conclusions

The proposed methodology presents a novel approach to knowledge graph embeddings by incorporating latent space distances and hierarchical clustering directly into the training process. This integration brings multiple advantages: first, the method is computationally efficient, scaling quadratically with the embedding dimensionality, which boosts scalability for larger graphs. Second, approximating the full likelihood eliminates the need for negative sampling, enabling it to capture the complete structure of the knowledge graph. Moreover, the embeddings effectively preserve hierarchical relationships, forming an ontology that mirrors the graph's inherent structures. Applied across six diverse knowledge graphs, this approach achieves competitive link prediction performance against numerous baseline models, even when using low-dimensional embeddings whereas the inferred hierarchy can be used as a data-driven approach to learn taxonomic representation in knowledge systems.

4.1 Limitations and broader impact

The inference procedures are subject to local minima and are not guaranteed to identify the optimal representation of knowledge graphs. Furthermore, the learning of hierarchies by use of clustering is NP-hard and prone to local minima issues. Whereas the full-likelihood approximation circumvents the need for negative sampling it still relies on the accuracy of the approximation which especially for low-dimensional embeddings can be coarse. Care has to be taken when using the model structure to infer missing or unobserved relations as the link prediction in some cases is far from perfect and may result in wrong conclusions in regard to entity-relationships. However, the approach can be used to efficiently probe facts unobserved to be verified or dismissed. Whereas we considered the widely used and well-established RESCAL framework, the approach readily extends to existing distance-based knowledge graph embedding procedures. Future work should investigate if other knowledge representation modeling procedures than RESCAL can be reformulated in terms of distance models amenable to similar scalable hierarchical approximations as the proposed HMSKGE.

References

Mehdi Ali, Max Berrendorf, Charles Tapley Hoyt, Laurent Vermue, Mikhail Galkin, Sahand Sharifzadeh, Asja Fischer, Volker Tresp, and Jens Lehmann. Bringing light into the dark: A large-scale evaluation of knowledge graph embedding models under a unified framework. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):8825–8845, 2021.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia: A nucleus for a web of open data. In *international semantic web conference*, pp. 722–735. Springer, 2007.

Ivana Balažević, Carl Allen, and Timothy M Hospedales. Tucker: Tensor factorization for knowledge graph completion. *arXiv preprint arXiv:1901.09590*, 2019.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pp. 1247–1250, 2008.

Antoine Bordes, Jason Weston, Ronan Collobert, and Yoshua Bengio. Learning structured embeddings of knowledge bases. In *Proceedings of the AAAI conference on artificial intelligence*, volume 25, pp. 301–306, 2011.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

Jiahang Cao, Jinyuan Fang, Zaiqiao Meng, and Shangsong Liang. Knowledge graph embedding: A survey from the perspective of representation spaces. *ACM Computing Surveys*, 56(6):1–42, 2024.

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam Hruschka, and Tom Mitchell. Toward an architecture for never-ending language learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 24, pp. 1306–1313, 2010.

Stefan Decker, Sergey Melnik, Frank Van Harmelen, Dieter Fensel, Michel Klein, Jeen Broekstra, Michael Erdmann, and Ian Horrocks. The semantic web: The roles of xml and rdf. *IEEE Internet computing*, 4 (5):63–73, 2000.

Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. Convolutional 2d knowledge graph embeddings. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 601–610, 2014.

Google. Introducing knowledge graph: Things, not strings, May 2012. URL `https://blog.google/products/search/introducing-knowledge-graph-things-not/`. Accessed: January 14, 2024.

Mark S Handcock, Adrian E Raftery, and Jeremy M Tantrum. Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(2):301–354, 2007.

Tue Herlau, Mikkel N Schmidt, and Morten Mørup. Infinite-degree-corrected stochastic block model. *Physical review E*, 90(3):032819, 2014.

Peter D Hoff, Adrian E Raftery, and Mark S Handcock. Latent space approaches to social network analysis. *Journal of the american Statistical association*, 97(460):1090–1098, 2002.

Charles Tapley Hoyt, Max Berrendorf, Mikhail Galkin, Volker Tresp, and Benjamin M Gyori. A unified framework for rank-based evaluation metrics for link prediction in knowledge graphs. *arXiv preprint arXiv:2203.07544*, 2022.

Yi Huang, Volker Tresp, Maximilian Nickel, Achim Rettinger, and Hans-Peter Kriegel. A scalable approach for statistical learning in semantic graphs. *Semantic Web*, 5(1):5–22, 2014.

Pengcheng Jiang, Shivam Agarwal, Bowen Jin, Xuan Wang, Jimeng Sun, and Jiawei Han. Text-augmented open knowledge graph completion via pre-trained language models. *arXiv preprint arXiv:2305.15597*, 2023.

Xueyan Jiang, Volker Tresp, Yi Huang, and Maximilian Nickel. Link prediction in multi-relational graphs using additive models. *SeRSy*, 919(2012):1–12, 2012.

Brian Karrer and Mark EJ Newman. Stochastic blockmodels and community structure in networks. *Physical review E*, 83(1):016107, 2011.

Charles Kemp, Joshua B Tenenbaum, Thomas L Griffiths, Takeshi Yamada, and Naonori Ueda. Learning systems of concepts with an infinite relational model. In *AAAI*, volume 3, pp. 5, 2006.

Tamara G Kolda, Brett W Bader, and Joseph P Kenny. Higher-order web link analysis using multilinear algebra. In *Fifth IEEE International Conference on Data Mining (ICDM'05)*, pp. 8–pp. IEEE, 2005.

Xiang Kong, Xianyang Chen, and Eduard Hovy. Decompressing knowledge graph representations for link prediction. *arXiv preprint arXiv:1911.04053*, 2019.

Pavel N Krivitsky, Mark S Handcock, Adrian E Raftery, and Peter D Hoff. Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. *Social networks*, 31(3):204–213, 2009.

Denis Krompaß, Maximilian Nickel, Xueyan Jiang, and Volker Tresp. Non-negative tensor factorization with rescal. In *Tensor Methods for Machine Learning, ECML workshop*, pp. 1–10, 2013.

Ni Lao and William W Cohen. Relational retrieval using a combination of path-constrained random walks. *Machine learning*, 81:53–67, 2010.

Garima Mishra, Anish R Khobragade, and Shashikant Ghumbre. Analysis of negative sampling methods for knowledge graph embedding. In *2023 7th International Conference On Computing, Communication, Control And Automation (ICCUBEA)*, pp. 1–5. IEEE, 2023.

Stephen H. Muggleton. Inverse entailment and progol. *New Generation Computing*, 13:245–286, 1995.

Nikolaos Nakis, Abdulkadir Çelikkanat, Sune Lehmann, and Morten Mørup. A hierarchical block distance model for ultra low-dimensional graph representations. *IEEE Transactions on Knowledge and Data Engineering*, 2023.

Mark EJ Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.

Maximilian Nickel and Volker Tresp. Learning taxonomies from multi-relational data via hierarchical link-based clustering. In *Learning Semantics. Workshop at NIPS*, volume 11, 2011.

Maximilian Nickel and Volker Tresp. Logistic tensor factorization for multi-relational data. *arXiv preprint arXiv:1306.2084*, 2013.

Maximilian Nickel, Volker Tresp, Hans-Peter Kriegel, et al. A three-way model for collective learning on multi-relational data. In *Icml*, volume 11, pp. 3104482–3104584, 2011.

Maximilian Nickel, Xueyan Jiang, and Volker Tresp. Reducing the rank in relational factorization models by including observable patterns. *Advances in Neural Information Processing Systems*, 27, 2014.

Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, 104(1):11–33, 2015.

Marcin Pietrasik and Marek Reformat. A simple method for inducing class taxonomies in knowledge graphs. In *European semantic web conference*, pp. 53–68. Springer, 2020.

J. Ross Quinlan. Learning logical definitions from relations. *Machine learning*, 5(3):239–266, 1990.

Steffen Rendle and Lars Schmidt-Thieme. Pairwise interaction tensor factorization for personalized tag recommendation. In *Proceedings of the third ACM international conference on Web search and data mining*, pp. 81–90, 2010.

Sebastian Riedel, Limin Yao, Andrew McCallum, and Benjamin M Marlin. Relation extraction with matrix factorization and universal schemas. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies*, pp. 74–84, 2013.

Daniel M Roy, Charles Kemp, Vikash Mansinghka, and Joshua Tenenbaum. Learning annotated hierarchies from relational data. *Advances in neural information processing systems*, 19, 2006.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15*, pp. 593–607. Springer, 2018.

Richard Socher, Danqi Chen, Christopher D Manning, and Andrew Ng. Reasoning with neural tensor networks for knowledge base completion. *Advances in neural information processing systems*, 26, 2013.

Ran Song, Shizhu He, Shengxiang Gao, Li Cai, Kang Liu, Zhengtao Yu, and Jun Zhao. Multilingual knowledge graph completion from pretrained language models with knowledge constraints. *arXiv preprint arXiv:2406.18085*, 2024.

Giorgos Stamou and Alexandros Chortaras. Ontological query answering over semantic data. *Reasoning Web. Semantic Interoperability on the Web: 13th International Summer School 2017, London, UK, July 7-11, 2017, Tutorial Lectures 13*, pp. 29–63, 2017.

Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pp. 697–706, 2007.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. Rotate: Knowledge graph embedding by relational rotation in complex space. *arXiv preprint arXiv:1902.10197*, 2019.

Ilya Sutskever, Joshua Tenenbaum, and Russ R Salakhutdinov. Modelling relational data using bayesian clustered tensor factorization. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta (eds.), *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009.

You Can Teach, Daniel Ruffinelli, Samuel Broscheit, and Rainer Gemulla. You can teach an old dog new tricks! on training knowledge graph embeddings. In *International Conference on Learning Representations*, 2020. URL `https://api.semanticscholar.org/CorpusID:211241737`.

Komal K. Teru, Etienne Denis, and William L. Hamilton. Inductive relation prediction by subgraph reasoning, 2020.

Kristina Toutanova and Danqi Chen. Observed versus latent features for knowledge base and text inference. In *Proceedings of the 3rd workshop on continuous vector space models and their compositionality*, pp. 57–66, 2015.

Liang Wang, Wei Zhao, Zhuoyu Wei, and Jingming Liu. Simkgc: Simple contrastive knowledge graph completion with pre-trained language models. *arXiv preprint arXiv:2203.02167*, 2022.

Meihong Wang, Linling Qiu, and Xiaoli Wang. A survey on knowledge graph embeddings for link prediction. *Symmetry*, 13(3):485, 2021.

David Kofoed Wind and Morten Mørup. Link prediction in weighted networks. In *2012 IEEE International Workshop on Machine Learning for Signal Processing*, pp. 1–6. IEEE, 2012.

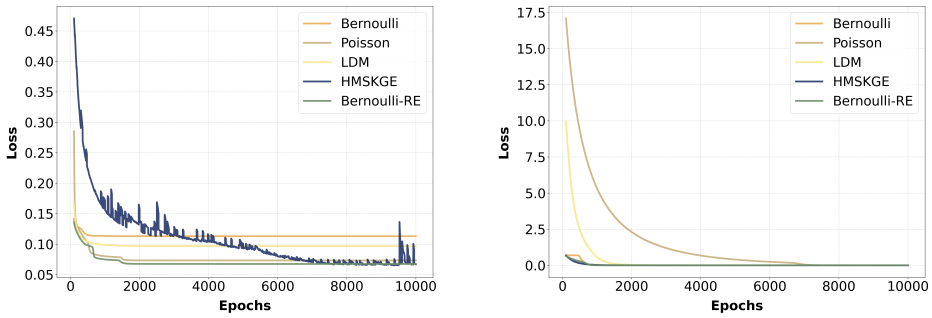
David H Wolpert. Stacked generalization. *Neural networks*, 5(2):241–259, 1992.

Liang Yao, Chengsheng Mao, and Yuan Luo. Kg-bert: Bert for knowledge graph completion. *arXiv preprint arXiv:1909.03193*, 2019.

Marinka Zitnik, Monica Agrawal, and Jure Leskovec. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*, 34(13):i457–i466, 2018.

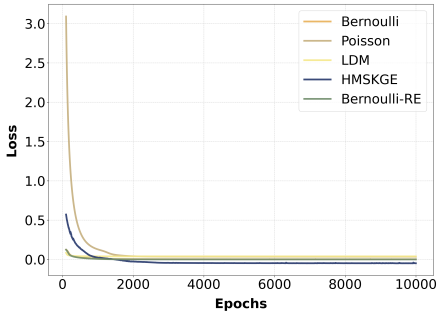
A Learning Curves

In Figure 4, the learning curves for all models applied to the *Kinships* and *WN18RR* knowledge graphs are presented. The results show that all models converge to a similar loss, indicating that *HMSKGE* effectively approximates the performance of the comparison models. We also observe that, in both cases, convergence is faster and smoother when using larger embedding dimensions.

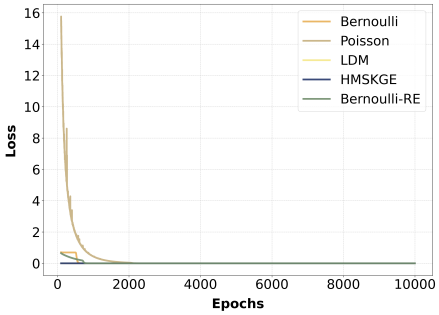


(a) *Kinships* (D = 2)

(b) *WN18RR* (D = 2)



(c) *Kinships* (D = 16)



(d) *WN18RR* (D = 16)

Figure 4: Learning curves for *Kinships* and *WN18RR* for dimensions 2 and 16. Plots start from epoch 100. HMSKGE losses are computed using the hierarchical approximation, whereas other models use the batched likelihood.

B Clustering

To assess the model's behavior in a controlled environment, HMSKGE was applied to an artificial knowledge graph comprising 30 entities and 3 relations, arranged into three predefined clusters. Each cluster interacted internally and with other clusters through specific relations. After randomizing the node order, the model successfully recovered the underlying clusters, as illustrated in Figure 5. This experiment demonstrates that HMSKGE can robustly infer cluster structure from relational patterns even when entity ordering provides no prior cues.

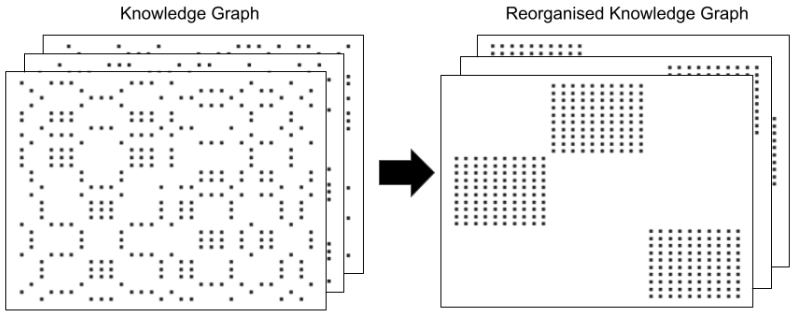


Figure 5: Reorganization of an artificial knowledge graph using HMSKGE, showing accurate recovery of three predefined clusters from randomized node order.