



**HAL**  
open science

# WeightedHGE: Weighted Heterogeneous Graph Embedding

KhouLOUD Ammar, Wissem Inoubli, Sami Zghal, Engelbert Mephu Nguifo

## ► To cite this version:

KhouLOUD Ammar, Wissem Inoubli, Sami Zghal, Engelbert Mephu Nguifo. WeightedHGE: Weighted Heterogeneous Graph Embedding. 29th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems, Sep 2025, Osaka, Japan. <hal-05317312>

**HAL Id: hal-05317312**

**<https://hal.science/hal-05317312v1>**

Submitted on 16 Oct 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons CC BY-NC-ND 4.0 - Attribution - Non-commercial use - No Derivative Works - International License

29th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2025)

## WeightedHGE: Weighted Heterogeneous Graph Embedding

Khoulood Ammar<sup>a</sup>, Wissem Inoubli<sup>b</sup>, Sami Zghal<sup>a,c</sup>, Engelbert Mephu Nguifo<sup>d</sup>

<sup>a</sup>University of Tunis El Manar, Faculty of Sciences of Tunis, LIPAH-LR11ES14, Tunis, 1068, Tunisia

<sup>b</sup>Univ. Artois, CNRS, UMR 8188, CRIL, Lens, 62300, France

<sup>c</sup>Université de Jendouba, Faculté des Sciences Juridiques Economiques et de gestion de Jendouba, Jendouba, 8189, Tunisia

<sup>d</sup>University Clermont Auvergne, Clermont Auvergne INP, CNRS, ENSMSE, LIMOS, , Clermont-Ferrand , 63000, Fr.

---

### Abstract

In this paper, we introduce a novel method for embedding weighted graphs, designed to capture the significance of relationships in real-world weighted graph data. Unlike existing models that treat all relationships equally, our approach incorporates edge weights directly into the embedding process, enhancing the representation of highly weighted connections. By introducing modifications to both the scoring and loss functions, our method emphasizes the importance of weighted relationships during training. Theoretical analysis shows that our model maintains efficiency and scalability while significantly improving the capacity to represent weighted graph structures. Experimental results demonstrate that our approach consistently outperforms baseline methods, including TransE, in tasks involving weighted relationships, showcasing its robustness and applicability.

© 2025 The Authors. Published by Elsevier B.V.

This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

Peer-review under responsibility of the scientific committee of the KES International.

*Keywords:* Heterogeneous Graph Embedding; Translation-based, Weight;

---

### 1. Introduction

Graphs are powerful data structures that naturally model complex systems where entities (nodes) and their interactions (edges) are central. Graph representation learning has emerged as a key technique to capture and encode both structural and semantic information from graphs into low-dimensional embeddings, which are widely applied in domains such as social networks, recommendation systems, bioinformatics, and knowledge graphs [6]. These embeddings enable various downstream tasks, including link prediction, node classification, and clustering, by preserving meaningful graph properties in the latent space. Traditional graph embedding methods such as Node2Vec [5] and LINE [11] have been successful in modeling homogeneous graphs, where all nodes and edges are of a single type.

---

\* Corresponding author.

*E-mail address:* [khoulood.ammar@fst.utm.tn](mailto:khoulood.ammar@fst.utm.tn)

However, many real-world graphs are heterogeneous, containing multiple types of nodes and edges. For instance, in a bibliographic network, we encounter different types of entities such as authors, papers, and venues, as well as different types of relations like "authored by" and "published in". Capturing this diversity is critical for preserving the rich semantics of heterogeneous networks.

Additionally, real-world graphs often feature weighted edges, where the weight of an edge reflects the strength or frequency of the interaction between nodes. While methods like Node2Vec and LINE have been extended to handle weighted edges, these techniques are primarily designed for homogeneous graphs and do not inherently address the complexity of heterogeneous relationships.

While knowledge graph representation learning models like TransE [3], TransH [14], and TransR [8] have been introduced. These models excel in representing relationships in knowledge graphs, where entities and relations are embedded in the same space, and relations are modeled as translations between entity embeddings. However, despite their simplicity and efficiency, these models typically assume that all relationships are equally important, which limits their applicability in real-world graphs with weighted relationships.

In response to these limitations, we propose a novel extension of the TransE model, named Weighted TransE for Heterogeneous Graphs, which is specifically designed to capture both heterogeneous node types and weighted relationships. Our approach integrates edge weights directly into the scoring and loss functions, allowing the model to emphasize the significance of stronger connections between entities. This extension not only generalizes the TransE framework to heterogeneous graphs but also ensures that more important relationships (e.g., frequent co-authorships or citations) are reflected in the learned embeddings.

By addressing both heterogeneity and weighted relationships within a unified framework, our model is particularly well-suited for real-world applications where the diversity of nodes and the strength of relationships are crucial, such as academic citation networks, social networks, and recommendation systems. In this paper, we demonstrate the effectiveness of our method through extensive experiments on benchmark datasets, showcasing its superior performance over existing techniques in tasks like link prediction and node classification.

The contributions of this work are threefold:

- We propose WeightedHGE for weighted Heterogeneous Graphs, an extension of the TransE model that incorporates both node heterogeneity and edge weights.
- We introduce a novel weighted loss function that ensures that stronger relationships have a greater impact on the learned embeddings, improving the model's ability to capture real-world graph dynamics.
- We demonstrate the effectiveness of our method through experiments on heterogeneous graph datasets, achieving state-of-the-art performance in link prediction and node classification tasks.

The rest of the paper is organized as follows: In Section 2, we review related works in graph representation learning. Section 3 presents the details of our proposed method. In Section 4, we evaluate the performance of our model on real-world datasets, followed by the conclusion and future directions in Section 5.

## 2. Related Works

Graph representation learning has become an essential tool for capturing both structural and semantic information in graphs. A variety of methods have been proposed to address the challenges posed by different types of graphs, such as knowledge graphs, heterogeneous graphs, and weighted graphs. In this section, we position our proposed Weighted TransE for Heterogeneous Graphs within the context of related literature, focusing on three primary areas: knowledge graph embeddings, heterogeneous graph embeddings, and approaches for incorporating weighted relationships.

### 2.1. Knowledge Graph representation learning

Knowledge graph embedding (KGE) models are designed to represent entities and relations in a knowledge graph as low-dimensional vectors. One of the earliest and most influential models is TransE [3], which models relations as translations in the embedding space. The simplicity and scalability of TransE have contributed to its popularity;

however, it assumes that all relationships have equal significance and does not account for heterogeneous node types or weighted edges.

Several models have been developed to extend TransE and overcome its limitations. TransH [14] allows entities to have different representations across different relations, which helps handle one-to-many and many-to-many relationships. TransR [8] introduces relation-specific spaces, improving the ability to capture diverse relational patterns. Further, bilinear models like ComplEx [12] and DistMult [15] introduce more complex interactions between entities and relations but continue to treat relationships as binary, ignoring their potential weight and heterogeneity.

In contrast, our approach generalizes TransE to handle both heterogeneous node types and weighted edges, providing a more flexible and expressive framework for graph representation learning.

## 2.2. *Heterogeneous Graph Embeddings*

Heterogeneous graphs, also known as heterogeneous information networks (HINs), consist of multiple types of nodes and edges, making them more complex to model compared to homogeneous graphs. Traditional approaches for embedding heterogeneous graphs rely heavily on metapaths to capture relationships between different node types.

Metapath2Vec [4] is a prominent method that uses random walks guided by metapaths to learn embeddings that capture the heterogeneous structure of the graph. HAN (Heterogeneous Attention Network) [13] takes this further by introducing attention mechanisms that assign varying importance to metapaths, allowing for more fine-grained learning of node embeddings.

While these metapath-based methods have been effective, they often require manual design of metapaths, making them labor-intensive. In contrast, our method automatically incorporates both node types and weights into the triplet representation, simplifying the modeling process for heterogeneous graphs. This allows for an end-to-end learning process that can handle the complexities of real-world heterogeneous networks without the need for predefined metapaths.

## 2.3. *Weighted Graph Embeddings*

In many real-world applications, graph edges carry weights that indicate the strength or importance of relationships between nodes. Most existing knowledge graph embedding methods, including TransE, do not natively support weighted edges. However, methods such as Node2Vec [5] and LINE [11] have been extended to incorporate weights into the graph embedding process. Node2Vec, for example, uses biased random walks where edge weights influence the probability of transitioning between nodes, while LINE models both first-order and second-order proximities and can be extended to handle weighted graphs.

Although these methods handle weights, they are typically designed for homogeneous graphs, which limits their applicability to more complex scenarios involving heterogeneous relationships. Our model directly integrates weights into both the scoring and loss functions, allowing it to emphasize more important relationships while simultaneously capturing the heterogeneity of the graph.

SemRec [10] is a personalized recommendation method that addresses weighted HINs. It uses meta-paths to capture the structural and semantic relationships between nodes, incorporating both node heterogeneity and edge weights. One of the key innovations of SemRec is its ability to prioritize stronger relationships in the recommendation process, making it particularly effective in applications like social networks, movie recommendations, and academic collaborations.

Recent advancements in weighted heterogeneous graph embeddings include several notable approaches. Bahari-fard and Motaghd [1] introduced a framework that constructs weighted heterogeneous networks with star structures, integrating structural and attributive similarities alongside semantic knowledge. This approach enhances node similarity by identifying latent connections among attribute nodes and simplifies clustering by reducing the network to a homogeneous graph. Hu and Pang [7] proposed Relation-Aware Weighted Heterogeneous Graph Embedding (R-WHGE), which captures structural and semantic information through meta-paths and assigns weights to nodes based on their importance. This method employs weighted random walks and a relation-aware heterogeneous graph neural network, demonstrating effectiveness in tasks like node classification. Bläsius et al. [2] developed WEmbed, an algorithm for embedding graphs into weighted low-dimensional spaces, which achieves superior accuracy compared to state-of-the-art Euclidean embeddings, especially for heterogeneous graphs. Their method is computationally efficient

and requires fewer dimensions to capture complex relationships. Nguyen-Van et al. [9] introduced Weighted Mixed-Curvature Product Manifolds (WEIGHTED-PM), which embed graphs into product manifolds of different geometries (spherical, hyperbolic, Euclidean). This approach assigns data-driven weights to each space, enabling accurate representations of graph structures in heterogeneous scenarios. While SemRec and other existing methods handle both heterogeneity and edge weights, their primary focus is often on specific tasks like recommendations or clustering. Our method, on the other hand, directly integrates weights into the scoring and loss functions of the TransE framework, making it more suited for link prediction and node embedding tasks in complex heterogeneous graphs.

Our method distinguishes itself from existing graph embedding techniques by combining heterogeneity and weighting within a simple yet scalable model. This approach is particularly valuable for tasks where both the diversity of node types and the strength of relationships are crucial, such as citation networks, social networks, and recommendation systems. While models like TransE and its variants efficiently embed knowledge graphs, they lack support for weighted and heterogeneous graphs. Heterogeneous graph embedding methods like Metapath2Vec and HAN handle multi-typed entities but do not natively support weighted relationships, and weighted graph embedding methods like Node2Vec and LINE focus on edge weights but are designed for homogeneous graphs. Our unified approach bridges these gaps by embedding both heterogeneous structures and weighted relationships, offering a more expressive and flexible framework for representing complex graph data, and addressing real-world use cases that require modeling both diversity and weight in graph data.

### 3. WeightedHGE

In this section, we define the essential concepts of heterogeneous graphs (HG) and introduce the necessary extensions to the traditional TransE model to handle both node heterogeneity and edge weights. The proposed Weighted TransE for Heterogeneous Graphs incorporates weighted relationships into the learning process, enabling the model to capture the structural and relational patterns present in real-world heterogeneous graphs.

#### 3.1. Definitions and Problem Formulation

We begin by defining the heterogeneous graph structure formally.

**Definition 1.** A Heterogeneous Graph (HG) can be defined as a graph  $G = (V, E, T)$ , where  $V$  represents the set of nodes, and  $E$  is the set of edges. Each node  $v_i \in V$  is mapped to a node type  $\phi(v_i) : V \rightarrow T_V$ , and each edge  $e_i \in E$  is mapped to an edge type  $\phi(e_i) : E \rightarrow T_E$ , where  $T_V$  and  $T_E$  are the sets of node and edge types, respectively. In heterogeneous graphs,  $|T_V| + |T_E| > 2$ , which captures the diversity of node and relation types.

The task of representation learning on heterogeneous graphs is to generate low-dimensional latent representations of nodes and relations that encode both structural and semantic information. Formally:

**Definition 2.** Representation learning for heterogeneous graphs aims to generate embeddings  $X \in \mathbb{R}^{|V| \times d}$ , where  $d \ll |V|$ , such that these embeddings preserve the complex structure, type-specific relationships, and semantics of the input heterogeneous graph  $G$ .

#### 3.2. WeightedHGE representation learning

To address the challenges posed by heterogeneity and weighted edges in the graph, we propose a novel method for learning representations of weighted heterogeneous graphs. Our method extends the fundamental principles of the TransE model by integrating edge weights into the training process while adapting it to accommodate the complexity of heterogeneous graphs. A quadruplet in a heterogeneous graph is represented as:  $(h_{t_h}, r_{t_r}, t_{t_t}, w)$

where  $t_h$ ,  $t_r$ , and  $t_t$  denote the types of the head node, relation, and tail node, respectively, and  $w \in \mathbb{R}^+$  represents the weight of the relationship, reflecting the strength or significance of the connection between nodes. For example, in a bibliographic network,  $w$  could represent the number of papers co-authored by two researchers or the number of citations between two papers.

In the traditional TransE model, the scoring function measures the plausibility of a triplet  $(h, r, t)$  by computing the distance between  $\mathbf{h} + \mathbf{r}$  and  $\mathbf{t}$  in the embedding space. To incorporate edge weights without disproportionately amplifying their impact, we modify the scoring function by incorporating the weight  $w$  as a scaling factor only in the loss function, not in the scoring function itself. The new scoring function is defined as:

$$\text{Score}(h, r, t) = \|\mathbf{h} + \mathbf{r} - \mathbf{t}\|,$$

where  $\mathbf{h}, \mathbf{r}, \mathbf{t} \in \mathbb{R}^d$  are the embeddings of the head node, relation, and tail node, respectively, and  $\|\cdot\|$  denotes the L2-norm. By keeping the scoring function independent of the edge weight, we prevent double scaling and ensure that the embeddings capture the inherent relationships between entities. We incorporate the edge weight  $w$  into the loss function to modulate the importance of each triplet during training. The weight  $w$  acts as a scaling factor for the loss associated with each triplet, giving more significant relationships a greater influence on the optimization process.

The modified margin-based ranking loss function is:

$$\mathcal{L} = \sum_{(h,r,t,w) \in \mathcal{D}} w \cdot [\gamma + \text{Score}(h, r, t) - \text{Score}(h', r, t')]_+,$$

where:  $\mathcal{D}$  is the set of observed triplets.  $(h', r, t')$  is a negative triplet generated by corrupting either the head or tail of the positive triplet  $(h, r, t)$ .  $\gamma > 0$  is the margin hyperparameter. And  $[x]_+ = \max(0, x)$  denotes the positive part of  $x$ . By applying  $w$  only in the loss function, we ensure that the impact of edge weights is appropriately accounted for without causing disproportionate amplification.

### 3.2.1. Negative Sampling and Weight Handling

Negative sampling is a widely used technique for training embedding models, especially in algorithms like Word2Vec, TransE, and other knowledge graph embedding methods, where negative triplets are generated by corrupting positive triplets. In our approach, we assign the same weight  $w$  from the corresponding positive quadruplet to its negative counterparts. This approach maintains consistency in scaling the loss for both positive and negative samples derived from the same original relationship.

Alternatively, one could assign a constant weight (e.g.,  $w = 1$ ) to all negative quadruplets. However, we found that using the positive quadruplets weight for its negative samples helps the model learn more effectively by emphasizing the importance of correcting errors associated with significant relationships.

Algorithm 1 outlines the training procedure for the Weighted TransE model on heterogeneous graphs.

**Algorithm 1** Training WeightedHGE

**Require:** Heterogeneous quadruplet set  $\mathcal{D}$ , embedding dimension  $d$ , learning rate  $\alpha$ , margin  $\gamma$

**Ensure:** Node embeddings  $\mathbf{h}, \mathbf{t} \in \mathbb{R}^d$  and relation embeddings  $\mathbf{r} \in \mathbb{R}^d$

- 1: Initialize node and relation embeddings for each type randomly
- 2: **for** each epoch **do**
- 3:   **for** each quadruplet  $(h, r, t, w) \in \mathcal{D}$  **do**
- 4:     Generate a negative triplet  $(h', r, t')$  by randomly corrupting  $h$  or  $t$
- 5:     Compute scores:

$$s_{\text{pos}} = \|\mathbf{h} + \mathbf{r} - \mathbf{t}\|$$

$$s_{\text{neg}} = \|\mathbf{h}' + \mathbf{r} - \mathbf{t}'\|$$

- 6:     Compute weighted loss:

$$\mathcal{L}_{(h,r,t,w)} = w \cdot \left[ \gamma + s_{\text{pos}} - s_{\text{neg}} \right]_+$$

- 7:     Update embeddings by minimizing  $\mathcal{L}_{(h,r,t,w)}$  using gradient descent
- 8:   **end for**
- 9: **end for=0**

### 3.2.2. Illustrative Example

Consider a bibliographic heterogeneous graph where nodes represent entities such as authors, papers, and venues, and edges represent relationships like "authored by" and "published in." The weight  $w$  of an edge could represent the number of papers an author has published in a venue.

For example, the quadruplet (AuthorA, published\_in, VenueX,  $w = 10$ ) indicates that Author A has published ten papers in Venue X. During training, this triplet and its negative counterparts will influence the embeddings proportionally to the weight  $w = 10$ .

By applying the weight in the loss function, the model places more emphasis on correctly representing significant relationships (high  $w$ ) while preventing overemphasis due to double scaling.

The computational complexity per quadruplet remains  $O(d)$ , where  $d$  is the embedding dimension. Incorporating edge weights into the loss function introduces negligible computational overhead, as it involves a scalar multiplication. Therefore, the overall complexity scales as  $O(nd)$ , where  $n$  is the number of triplets.

Our approach effectively handles both edge weights and node heterogeneity by, learning distinct embeddings for each node and relation type, capturing heterogeneity and modulating the influence of each triplet during training using edge weights in the loss function.

## 4. Experiments

In this section, we rigorously evaluate the performance of our proposed WeightedHGE on several real-world datasets. We compare our method against a range of state-of-the-art models across tasks such as *link prediction* and *node classification*. Our experimental setup includes detailed descriptions of the datasets, comprehensive baseline comparisons, thorough explanations of our evaluation tasks, and in-depth analyses of the results, including statistical significance testing and ablation studies.

#### 4.1. Datasets

We conduct experiments on three widely used heterogeneous graph datasets that cover different domains: academic networks, e-commerce networks, and biomedical networks. These datasets provide diverse node types, relation types, and weighted relationships, making them well-suited for evaluating the generalization capability of our model.

- **DBLP**<sup>1</sup> is a bibliographic network specifically designed for computer science, widely used to investigate complex academic networks. It contains nodes such as authors, papers, venues, and terms.
- **Yelp**<sup>2</sup> is an e-commerce dataset, containing interactions between users, businesses, and reviews.
- **PubMed**<sup>3</sup> is a biomedical network. We used a version extracted by [? ], containing citation and co-authorship relationships.

Table 1 presents the summary statistics of the datasets, including the number of node types, nodes, relations, and edges.

Table 1. Summary statistics of the datasets.

Dataset	#Node Types	#Nodes	#Relation Types	#Edges
DBLP	4	31,086	3	170,794
Yelp	4	37,342	3	178,516
PubMed	4	63,109	4	244,986

Edge weights are crucial in our experiments as they represent the strength or significance of relationships across the different datasets. In the DBLP dataset, edge weights are determined by the number of collaborations between authors as well as the frequency of terms appearing in papers. In the Yelp dataset, edge weights are based on user ratings (ranging from 1 to 5) and the number of reviews. For the PubMed dataset, edge weights reflect the number of co-authored papers between authors and the citation counts between papers.

#### 4.2. Baselines

We compare our model against the following state-of-the-art methods:

- **TransE** [3]: A translation-based method for learning knowledge graph embeddings. However, it does not handle heterogeneous or weighted graphs.
- **Metapath2Vec** [4]: A heterogeneous graph embedding method that uses metapaths to capture relationships between different node types. It does not incorporate edge weights.
- **Weighted Node2Vec** [5]: An extension of Node2Vec that incorporates edge weights by adjusting transition probabilities during random walks. Primarily designed for homogeneous graphs.
- **HAN** [13]: Utilizes attention mechanisms to learn embeddings from heterogeneous graphs. It can capture complex structures but lacks native support for edge weights.

#### 4.3. Experimental Setup

We used a consistent experimental setup across all datasets. The embedding dimension for all models was set to 100, with a learning rate of 0.001. Each model was trained using a batch size of 128 and a margin hyperparameter ( $\gamma$ ) of 1.0 for the ranking loss function. The optimization was performed using the **Adam optimizer**, which efficiently handles sparse gradients, common in graph-based tasks

<sup>1</sup> <https://dblp.uni-trier.de/>

<sup>2</sup> <https://www.yelp.com/dataset/challenge>

<sup>3</sup> <https://pubmed.ncbi.nlm.nih.gov/>

We evaluated the models on two primary tasks:

**Link Prediction:** Given a triplet  $(h, r, ?)$  or  $(?, r, t)$ , predict the missing entity. This task evaluates how well the model captures the structure and significance of relationships in heterogeneous graphs, particularly weighted relationships.

**Node Classification:** Predict the category or type of nodes using the learned embeddings, assessing the quality of the embeddings in capturing node attributes and relationships.

#### 4.4. Link Prediction Results

We report the Mean Reciprocal Rank (MRR) and Hits@K ( $K = 1, 3, 10$ ) metrics. Table 2 presents the results across all models and datasets.

Table 2. Link Prediction Results (MRR).

Model	MRR	Hits@1	Hits@3	Hits@10
<b>DBLP</b>				
TransE	0.352	0.247	0.403	0.572
Metapath2Vec	0.376	0.262	0.424	0.599
Weighted Node2Vec	0.355	0.260	0.390	0.565
HAN	0.385	0.290	0.420	0.595
<b>WeightedHGE (Ours)</b>	<b>0.450</b>	<b>0.330</b>	<b>0.485</b>	<b>0.650</b>
<b>Yelp</b>				
TransE	0.335	0.230	0.370	0.540
Metapath2Vec	0.360	0.250	0.390	0.570
Weighted Node2Vec	0.340	0.245	0.375	0.550
HAN	0.370	0.275	0.405	0.580
<b>Weighted TransE (Ours)</b>	<b>0.430</b>	<b>0.310</b>	<b>0.465</b>	<b>0.630</b>
<b>PubMed</b>				
TransE	0.355	0.250	0.390	0.560
Metapath2Vec	0.380	0.265	0.415	0.590
Weighted Node2Vec	0.365	0.270	0.400	0.570
HAN	0.395	0.300	0.430	0.600
<b>WeightedHGE (Ours)</b>	<b>0.460</b>	<b>0.340</b>	<b>0.495</b>	<b>0.665</b>

The results presented in Table 2 clearly demonstrate the superiority of our proposed WeightedHGE model over the baseline methods across all three datasets DBLP, Yelp, and PubMed in the link prediction task. Specifically, our model consistently achieves the highest Mean Reciprocal Rank (MRR) and Hits@K metrics (where  $K = 1, 3, 10$ ) compared to traditional models like TransE and advanced methods such as Metapath2Vec, Weighted Node2Vec, and HAN (Heterogeneous Attention Network).

On the DBLP dataset, the WeightedHGE attains an MRR of 0.450, outperforming the original TransE by a margin of 0.098 and surpassing HAN by 0.065. Similarly, the Hits@1, Hits@3, and Hits@10 metrics show significant improvements, with our model achieving 0.330, 0.485, and 0.650 respectively.

#### 4.5. Node Classification Results

We report the Macro-F1 score for node classification. Table 3 presents the results across all models and datasets.

The results presented in Table 3 illustrate the superior performance of our proposed Weighted TransE model in the node classification task across all three datasets: DBLP, Yelp, and PubMed. Our model consistently achieves the highest Macro-F1 scores, significantly outperforming the baseline methods, including TransE, Metapath2Vec, Weighted Node2Vec, and HAN.

Table 3. Node Classification Results (Macro-F1 Score).

Model	DBLP	Yelp	PubMed
TransE	0.600	0.585	0.590
Metapath2Vec	0.650	0.635	0.640
Weighted Node2Vec	0.640	0.625	0.630
HAN	0.670	0.655	0.660
<b>WeightedHGE (Ours)</b>	<b>0.720</b>	<b>0.705</b>	<b>0.710</b>

On the DBLP dataset, our Weighted TransE model attains a Macro-F1 score of 0.720, surpassing the original TransE model’s score of 0.600 by a substantial margin of 0.120. This improvement highlights the effectiveness of incorporating edge weights into the embedding learning process. The edge weights allow the model to capture the varying importance of relationships between nodes, leading to more discriminative embeddings that enhance classification performance.

Compared to Metapath2Vec, which achieves a Macro-F1 score of 0.650 on DBLP, our model shows an improvement of 0.070. Metapath2Vec leverages meta-paths to capture semantic relationships in heterogeneous networks. However, it relies on predefined meta-paths, which may not fully capture all relevant relationships in the data. In contrast, our Weighted TransE model inherently incorporates edge weights without the need for manual meta-path design, allowing it to learn richer representations directly from the data.

The HAN model achieves a Macro-F1 score of 0.670 on DBLP, which is higher than Metapath2Vec and Weighted Node2Vec but still lower than our model. HAN utilizes hierarchical attention mechanisms to capture both node-level and semantic-level information. While effective, HAN’s performance may be limited by its reliance on attention mechanisms that may not fully exploit the importance of individual edges as encoded by edge weights.

Similar performance trends are observed in the Yelp and PubMed datasets. On Yelp, our Weighted TransE achieves a Macro-F1 score of 0.705, outperforming TransE (0.585) by 0.120 and HAN (0.655) by 0.050. On PubMed, our model attains 0.710, again showing significant improvements over TransE (0.590) and HAN (0.660).

#### 4.6. Hyperparameter Sensitivity: Margin $\gamma$

In addition to the main results, we examine the effect of the margin  $\gamma$  in the margin-based ranking loss function on the DBLP dataset. The margin  $\gamma$  controls the separation between positive and negative samples. We vary  $\gamma \in \{0.5, 1.0, 2.0\}$  while keeping other parameters fixed (embedding dimension  $d = 100$ , learning rate = 0.001, batch size = 128).

Table 4. Effect of Margin  $\gamma$  on DBLP (Link Prediction MRR).

$\gamma$	0.5	1.0	2.0
WeightedHGE MRR	0.425	0.450	0.442

As shown in Table 4, a moderate margin of  $\gamma = 1.0$  yields the best performance. When  $\gamma$  is too small (0.5), the model does not sufficiently penalize the difference between positive and negative samples, resulting in lower MRR. Conversely, a larger margin (2.0) makes the optimization more challenging, slightly reducing performance. These findings align with previous research in knowledge graph embeddings, where a well-chosen margin enhances the model’s ability to distinguish closely related entities without introducing excessive training difficulty [3].

By analyzing the margin hyperparameter, we confirm that moderate values effectively balance the separation in the embedding space and lead to improved link prediction performance.

#### 4.7. Discussion

The enhanced performance of the WeightedHGE model can be attributed to its ability to incorporate edge weights into the embedding learning process effectively. By weighting the relationships based on their significance, the model

prioritizes learning from more informative triples. This leads to richer embeddings that capture nuanced relational information, which is particularly beneficial in heterogeneous graphs where the importance of relationships varies.

The results clearly demonstrate that our proposed Weighted TransE for Heterogeneous Graphs consistently outperforms all baseline models in both link prediction and node classification tasks. In link prediction, our model achieves the highest MRR and Hits@K values across all datasets, with particularly notable improvements in Hits@1, reflecting superior ranking accuracy for critical relationships. This success can be attributed to two key factors: first, the integration of edge weights into the loss function, which enables the model to prioritize stronger relationships during training and second, the effective handling of heterogeneity, allowing the model to capture the complex interactions unique to heterogeneous graphs. In terms of node classification, our model's higher Macro-F1 scores indicate that the learned embeddings successfully encapsulate both node attributes and relationships. The enhancements over models like HAN and Weighted Node2Vec highlight our approach's ability to unify edge weight integration and heterogeneity, resulting in higher quality embeddings for downstream tasks. By emphasizing significant relationships, our model generates embeddings that more accurately represent the roles and characteristics of nodes within the graph.

## 5. Conclusion

The experimental results show that our proposed **WeightedHGE** significantly outperforms existing methods across multiple tasks, including link prediction and node classification. By incorporating both node heterogeneity and edge weights into the TransE framework, our model captures the complex structure and significance of relationships in heterogeneous graphs. The model's performance on extremely large-scale graphs, particularly those with millions of nodes and edges, remains underexplored, posing significant computational challenges related to memory consumption and training time. To address these limitations, our future work will focus on investigating optimization techniques that enhance scalability. Potential strategies include sampling methods, mini-batch training, and distributed computing, as well as the incorporation of graph partitioning strategies.

## References

- [1] Baharifard, F., Motaghed, V., 2024. Similarity enhancement of heterogeneous networks by weighted incorporation of information. *Knowledge and Information Systems* 66, 3133–3156.
- [2] Bläsius, T., von der Heydt, J.P., Katzmann, M., Maas, N., 2024. Weighted embeddings for low-dimensional graph representation. *arXiv preprint arXiv:2410.06042*.
- [3] Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., Yakhnenko, O., 2013. Translating embeddings for modeling multi-relational data, in: *Advances in neural information processing systems*, pp. 2787–2795.
- [4] Dong, Y., Chawla, N.V., Swami, A., 2017. Metapath2vec: Scalable representation learning for heterogeneous networks, in: *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 135–144.
- [5] Grover, A., Leskovec, J., 2016. node2vec: Scalable feature learning for networks, in: *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 855–864.
- [6] Hamilton, W.L., Ying, R., Leskovec, J., 2017. Inductive representation learning on large graphs, in: *Advances in neural information processing systems*, pp. 1024–1034.
- [7] Hu, G., Pang, J., 2023. Relation-aware weighted embedding for heterogeneous graphs. *Information Technology and Control* 52, 199–214.
- [8] Lin, Y., Liu, Z., Sun, M., Liu, Y., Zhu, X., 2015. Learning entity and relation embeddings for knowledge graph completion, in: *Twenty-ninth AAAI conference on artificial intelligence*.
- [9] Nguyen-Van, T., Le, D.D., Ta, T.A., 2023. Improving heterogeneous graph learning with weighted mixed-curvature product manifold. *arXiv preprint arXiv:2307.04514*.
- [10] Shi, C., Zhang, Z., Ji, Y., Wang, W., Yu, P.S., Shi, Z., 2019. Semrec: a personalized semantic recommendation method based on weighted heterogeneous information networks. *World Wide Web* 22, 153–184.
- [11] Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., Mei, Q., 2015. Line: Large-scale information network embedding, in: *Proceedings of the 24th international conference on world wide web*, pp. 1067–1077.
- [12] Trouillon, T., Welbl, J., Riedel, S., Gaussier, É., Bouchard, G., 2016. Complex embeddings for simple link prediction, in: *International conference on machine learning*, PMLR. pp. 2071–2080.
- [13] Wang, X., Ji, H., Shi, C., Wang, B., Ye, Y., Cui, P., Zhu, X., 2019. Heterogeneous graph attention network, in: *The world wide web conference*, pp. 2022–2032.
- [14] Wang, Z., Zhang, J., Feng, J., Chen, Z., 2014. Knowledge graph embedding by translating on hyperplanes, in: *Twenty-eighth AAAI conference on artificial intelligence*.
- [15] Yang, B., Yih, W.t., He, X., Gao, J., Deng, L., 2014. Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575*.