

# Composition of Heterogeneous Node Embeddings - Unlocking the Power of Heterogeneous Graph Representation

Silvio Fernando Angonese<sup>1</sup>, Renata Galante<sup>1</sup>

<sup>1</sup>Institute of Informatics, Federal University of Rio Grande do Sul (UFRGS)  
P.O. Box 15,064 – ZIP Code 91501-970 – Porto Alegre, RS, Brazil

{sfangonese, galante}@inf.ufrgs.br

**Abstract.** *Heterogeneous graphs have high representation power, which can be maximized through node embeddings. Important embedding approaches are based on node features and node metapaths, applied individually. This paper proposes the creation of heterogeneous composition node embeddings, which are based on local node features, features from node neighbors, and node metapaths. This results in two types of composition embeddings: Features + Metapaths and Aggregated + Metapaths. Experiments have demonstrated superior performance compared to the baseline. In the experiments, our composition Aggregated Features + Metapaths embedding achieved a Micro-F1 score of 65.89% compared to 61.53% from the baseline, highlighting its effectiveness. Additionally, this paper also evaluates alternative models with these embedding compositions that outperform the state-of-the-art approach.*

## 1. Introduction

Heterogeneous graphs are important data structures used to represent both simple and complex data-driven applications. They provide an excellent solution for representation learning and as input datasets in various types of downstream applications. A critical research question in this context is how we can enhance the representational power of heterogeneous graphs and their components. Successfully addressing this question could directly benefit applications that utilize these graphs, potentially leading to improved performance.

One promising approach to enhancing this representational power is through the use of node embeddings. Node embeddings are vector representations of nodes in graphs that capture their features and relationships, allowing Deep Learning and Machine Learning algorithms to operate efficiently [Wang et al. 2023]. In heterogeneous graphs, metapaths represent sequences of relationships connecting different types of nodes, providing a means to explore and integrate complex structural information. Combining node embeddings with metapaths can significantly improve the semantic representation and performance of graph-based applications.

Some work [Hamilton et al. 2017, Ying et al. 2018] introduces the generation of embeddings based on the neighboring nodes, increasing their expressiveness and recommendation performance. The work [Wang et al. 2023] proposes a new approach for generating embeddings from metapaths, capturing semantics based on node relationships. MAGNN [Fu et al. 2020] is a neural network designed to incorporate node content along metapaths in heterogeneous graphs, and aims to improve graph embedding by aggregating information from multiple metapaths to enrich node representation. Due to its relevance to

our research and its status as a state-of-the-art method, MAGNN was chosen as the baseline for our paper. However, none of the related work addresses the use of heterogeneous graphs with the composition of embeddings to increase the semantic representativeness of the nodes. Since nodes have embeddings composed of local information, their neighbors, and their relationships with different types of nodes, this significantly enhances their representational power.

In our previous work AGHE [Angonese and Galante 2024], we proposed an approach that creates a heterogeneous graph with embeddings generated from various heterogeneous node features such as texts, images, and subgraphs. Thus, each node not only has features but also individual embeddings for features and metapaths, and the merge of features with metapaths. The gap in this work is the lack of a definition for composition embeddings and the possibility of creating compositions based on features and metapaths.

This paper aims to specify a mechanism for aggregating semantics for nodes in heterogeneous graphs, proposing a novel type of node embeddings composed of aggregated features from neighbors and metapaths. Through the aggregation of features, the target node merges its features with those received from its neighbors, and the metapaths representing the relationships between the local type node and other types of nodes from its neighbors. It is the foundation for creating heterogeneous embeddings based on aggregated features from neighbors and metapaths. Experiments were conducted to compare the performance of classifiers in the Node Classification task with the MAGNN baseline. Our composition Aggregated Features + Metapaths embedding achieved a Micro-F1 score of 65.89% compared to 61.53% from the baseline, highlighting its effectiveness. The results have demonstrated the robustness and effectiveness of our proposal to use the embedding compositions to enhance node representation.

The remainder of this paper is organized as follows: Section 2 conceptualizes the background techniques applied in this paper. Section 3 reviews related work. Section 4 presents the AGHE approach as the base for the graph creation. Section 5 describes the core foundation of the proposed embeddings. Section 6 presents the experiments and results achieved, while Section 7 discusses the conclusions and future work.

## 2. Conceptualization

In this section, we present several important techniques closely related to the subject matter of this paper, providing crucial support and context. **Heterogeneous Graph** can have nodes and edges of different types, e.g., the graph encoding the relationship between the Movie and its Director and Actor [Fu et al. 2020]. The challenge of the heterogeneous graph representation learning is to figure out the information of nodes from it and their neighborhoods, which makes the aggregated embedding more powerful [Zhang et al. 2019, Sun et al. 2020, Jin et al. 2021]. The central problem in Deep Learning on graphs is finding a way to incorporate information about graph structure into Deep Learning models. From this perspective, the challenge is that there is no straightforward way to encode this high-dimensional, non-Euclidean information about graph structure into a feature vector [Hamilton et al. 2017].

**Graph Node Embeddings** is the node representation, aiming to learn a function  $f(x) : \mathcal{V} \rightarrow \mathcal{R}^d$  that embeds the nodes  $v$  into a low-dimensional Euclidean space with  $d \ll |\mathcal{V}|$  [Wang et al. 2023]. Thus, graph embedding is the transforma-

tion of property graphs to a vector or a set of vector spaces. Embedding should capture the graph topology, node features, node-to-node relationship, and other relevant information about graphs, subgraphs, and nodes. The similarity of embedding between nodes indicates their similarity in the network, i.e., both nodes are close to each other, connected or not by an edge, potentially used for any kind of prediction [Hamilton et al. 2017, Santana and Ribeiro 2023].

**Metapaths** are sequences of node types that define specific paths through heterogeneous graphs, capturing the semantics and structural correlations between different types of nodes [Sun and Han 2012]. Metapaths enable the analysis of complex relationships and interactions within the graph by providing a structured way to traverse and connect different types of entities. This concept involves creating graph embeddings based on random walks that explore the heterogeneous neighborhood of a node, considering the type constraints imposed by the metapath. Skip-Gram and Node2Vec models are employed to maximize the probability of preserving both the structural and semantic properties of the graph, thus enabling the learning of desirable node representations. By leveraging metapaths, these models can capture rich contextual information and improve the accuracy and interpretability of the resulting embeddings, making them highly effective for tasks such as Link Prediction, Node Classification, and Clustering in heterogeneous graphs [Dong et al. 2017].

**Node Classification** is the important RecSys task, aiming to predict the label  $y_v$  which could be a type, category, or attribute associated with all the nodes  $v \in \mathcal{V}$  when we are only given the true labels on a training set of nodes  $\mathcal{V}_{train} \subseteq \mathcal{V}$ . Thus, can make predictions  $\mathcal{Z}$  for each of the nodes by applying a shared function  $f$  to each of the latent vectors  $h$ , classifying nodes based on their features, as  $\mathcal{Z}_i = f(h_i)$  [Hamilton et al. 2017]. Examples of Node Classification could include determining the genre of movies based on their features and relationships. Node Classification models can exploit not only node features, but also the concept that nodes with similar local neighborhood structures tend to have similar labels. Additionally, they leverage the heterophily concept, which suggests that nodes are more likely to connect to others with different labels and types. Another interesting approach for Node Classification is based on vector embeddings, which prove extremely useful as feature inputs. The basic idea is to use information about the neighborhood of the node in a vector embedding, which serves as the representation of nodes [Hamilton et al. 2017].

### 3. Related Work

The heterogeneous graph can be traced back to generate data embedding from node features based on random walks approach citing Representation Learning on Graphs [Hamilton et al. 2017, Ying et al. 2018] improving the node expressivity. More close to the aims of our proposal is [Zhang et al. 2019] which defines of Heterogeneous Graph Neural Network with the processing of embedding. The survey Graph Neural Networks in RecSys [Wu et al. 2022] shows GNNs have been widely used in downstream applications essentially because graph structure and GNN have superiority in graph representation learning, citing GraphSAGE [Hamilton et al. 2017] as an important work regarding generating node embedding from node feature information.

MetaPath2Vec [Dong et al. 2017] is another crucial technique of this research due

to its ability to capture the structure of heterogeneous graph, guiding random walks to generate sequences of heterogeneous nodes with rich semantics. The Metapath Aggregated Graph Neural Network (MAGNN) is an approach for heterogeneous graph embedding, aiming to comprehensively consider the information present in heterogeneous graphs, based on the Intra-Metapath aggregation extracting and combining information from metapath instances connecting nodes with their neighbors. This node representation is used as input for an external classifier, such as SVM, to perform the Node Classification task [Fu et al. 2020]. This work is important and closely related to our approach, where the main distinction lies in the method used by MAGNN to aggregate node information from each node reached by metapaths, while our paper focuses on embedding compositions, capturing semantics from local and neighboring node information, and metapaths.

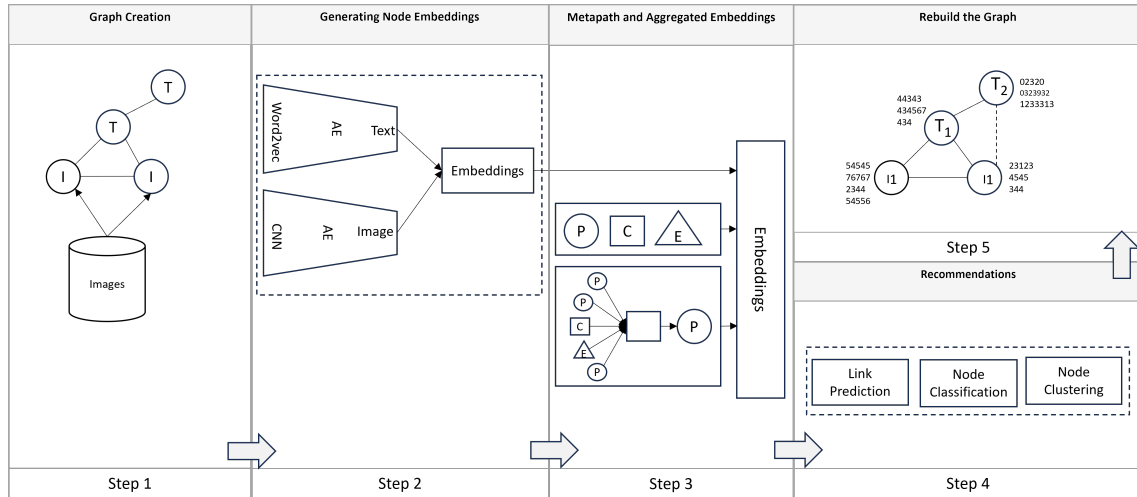
Heterogeneous embeddings and their composition can enhance the data quality of heterogeneous graphs, according to the research Approach for Generating Enhanced Heterogeneous Embeddings from Heterogeneous Graphs (AGHE) presented by [Angonese and Galante 2024]. Our paper expands on this concept by introducing a new composition with aggregated features from the node neighbors and metapaths from the target node. MAGNN was selected as a baseline because their approach is based on Intra-Metapaths, which are close to our embedding compositions, supporting direct comparisons. They used the public IMDb Movies India dataset, thus allowing us to reproduce the experiments using the same dataset. This paper demonstrates that the composition of features and metapaths embeddings outperforms both simple features and metapaths individually, as well as addressing the gap in MAGNN research, which does not use the composition of local and neighbor node features with their respective metapaths.

#### **4. AGHE - Approach for Generating Enhanced Heterogeneous Embeddings from Heterogeneous Graphs**

The *Approach for Generating Enhanced Heterogeneous Embeddings from Heterogeneous Graphs* (AGHE) [Angonese and Galante 2024] shown in Fig. 1, generates heterogeneous embeddings through the processing of texts, images, and subgraphs represented in the nodes of heterogeneous graphs, such as the following steps. In the next section, we define the compositions of embeddings introduced in this paper, which are used in Steps 2 and 3.

1. *Graph Creation* - generates the heterogeneous graph along with all its components, such as nodes, edges, and node features. This step is critical because all other steps and results depend on it;
2. *Generating Text Node Embeddings* - is the process of creating node embeddings from their corresponding node features or extracted from images embedded in the nodes;
3. *Metapath and Aggregated Node Embeddings* - generates of aggregated feature embeddings using the random walks approach, representing the business rules through the relationships among the nodes, followed by the generation of their embeddings using the MetaPath2Vec algorithm;
4. *Graph Enhancement with RecSys tasks* - represents the experiments conducted in this paper, aiming to predict the type of nodes, predict some links, and cluster the nodes based on the heterogeneous graph generated in the before steps;

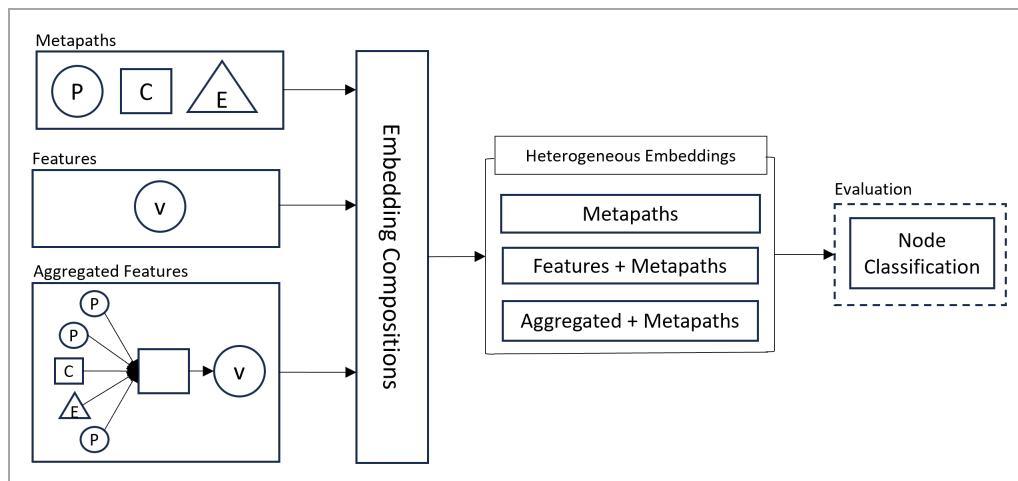
5. *Rebuilding the Graph* - involves incorporating the generated embeddings and predictions saved into the graph nodes.



**Figura 1. Steps of AGHE - Approach for Generating Enhanced Heterogeneous Embeddings from Heterogeneous Graphs.**

### 5. Composition of Heterogeneous Node Embeddings

This section presents the core foundation proposal of this paper for creating compositions of heterogeneous node embeddings based on node features and metapaths. We propose the creation of three embeddings composition based on node metapaths (Metapath), node features (Features + Metapaths), and aggregation of neighboring node features (Aggregated + Metapaths), and shown in Fig. 2.



**Figura 2. Embedding compositions proposal.**

#### 5.1. Metapaths Embedding

Metapaths are sequences of node types that define specific paths through heterogeneous graphs, and MetaPath2Vec is an algorithm that leverages the concept of metapaths, capturing complex structural and semantic relationships between different types of nodes in the

graph. A metapath embedding is generated by traversing the predefined metapaths and incorporating the information into the target node. Let  $\mathcal{HG} = (\mathcal{V}, \mathcal{E})$  be a heterogeneous graph where  $\mathcal{V}$  is the set of nodes and  $\mathcal{E}$  is the set of edges. A metapath  $\mathcal{M}$  is a sequence of node types and edges denoted as:

$$\mathcal{M} = (V_1 \xrightarrow{E_1} V_2 \xrightarrow{E_2} \dots \xrightarrow{E_{m-1}} V_m), \quad (1)$$

where  $V_i$  represents the node type and  $E_i$  represents the edge. For a target node  $v \in \mathcal{V}$ , the metapath embedding  $\mathbf{h}_v^M$  is generated by aggregating the information from nodes reached by traversing the metapath  $\mathcal{M}$  starting from  $v$ , formally defined as:

$$\mathbf{h}_v^M = \text{Aggregate}(\{f(u) \mid u \in \text{Reachable}(v, \mathcal{M})\}), \quad (2)$$

where  $\text{Reachable}(v, \mathcal{M})$  is the set of nodes that can be reached from  $v$  by following the metapath  $\mathcal{M}$ ,  $f(u)$  is a function that extracts the embedding of node  $u$ , and  $\text{Aggregate}$  is a function that combines these embeddings into a single embedding for the central node  $v$ .

## 5.2. Features + Metapaths Embedding

This embedding composition is the node representation with local information and its relationships with its neighbors. It is composed of local node features with metapath embeddings, capturing the semantics of the relationships with its neighbor nodes. For a node  $v \in \mathcal{V}$ , let  $\mathbf{x}_v$  be the feature vector representing the local information of node  $v$ . The embedding composition  $\mathbf{z}_v$  of node  $v$  is then defined as the concatenation of its local feature vector  $\mathbf{x}_v$  and its metapath embedding  $\mathbf{h}_v^M$ , which is denoted as:

$$\mathbf{z}_v = \mathbf{x}_v \parallel \mathbf{h}_v^M, \quad (3)$$

where  $\mathbf{h}_v^M$  is defined in Equation 2 and  $\parallel$  denotes the concatenation operation.

## 5.3. Aggregated + Metapaths Embedding

It is composed of aggregated node features from both local information and information from its neighbors through the random walk approach, which is then fused with metapath embeddings. This process captures the semantics of the relationships between the neighbor nodes, resulting in a node representation that incorporates local and neighbor semantics. For a node  $v \in \mathcal{V}$ , let  $\mathbf{a}_v$  be the aggregated feature vector that includes both the local information of node  $v$  and the information from its neighbors. Thus, this composition can be formally denoted as:

$$\mathbf{a}_v = \text{Aggregate}(\{f(u) \mid u \in \text{Neighbors}(v) \cup \{v\}\}), \quad (4)$$

where  $\text{Neighbors}(v)$  is the set of neighbor nodes of  $v$ ,  $f(u)$  is a function that extracts the feature vector of node  $u$ , and  $\text{Aggregate}$  is a function that combines these feature vectors. The embedding composition  $\mathbf{z}_v$  of node  $v$  is then defined as the concatenation of its aggregated feature vector  $\mathbf{a}_v$  and its metapath embedding  $\mathbf{h}_v^M$  from Equation (2):

$$\mathbf{z}_v = \mathbf{a}_v \parallel \mathbf{h}_v^M, \quad (5)$$

where  $\mathbf{h}_v^M$  is defined in Equation (2) and  $\parallel$  denotes the concatenation operation.

## 6. Experiments

We present the experiments conducted to investigate the effectiveness of the proposed embedding compositions based on features and metapaths. We also aim to evaluate the performance of a set of classifiers beyond those used in the baseline, providing an extensive comparison. Additionally, we compare the Hold-Out method used in the baseline with the proposed Cross-Validation method to achieve a more comprehensive and robust evaluation. These analyses support an understanding of how well embedding compositions enhance node representation and Node Classification task results in heterogeneous graphs. Consequently, we can demonstrate the strengths and limitations of each method in handling complex data structures, providing insights into their practical applicability in real-world scenarios.

### 6.1. Experiment Setup

We outline the experimental setup detailing the procedures and methodologies employed to investigate the performance metrics of the proposed embedding compositions. Comprehensive experiments using the IMDb India Movies <sup>1</sup> **Dataset** were conducted for the Node Classification task. IMDb is a collection of Indian movies with some data such as titles, genres, ratings, and cast members. In the data pre-processing, we selected only the genres *Action*, *Comedy*, and *Drama*, which are exactly used in the MAGNN baseline. Additionally, we selected movies with positive values for Year, Votes, and Rating, and containing Director information and at least one Actor. Thus, creating a subset with no duplicated movies to be used in all experiments of this paper.

Metapath Aggregated Graph Neural Network for Heterogeneous Graph Embedding (MAGNN) [Fu et al. 2020] is an important state-of-the-art research framework, and it serves as the **baseline for this paper**. MAGNN uses Intra-Metapaths embedding which is the aggregation of structural and semantic information from the target nodes and their neighbors along a specific metapath in a heterogeneous graph, applying the external SVM classifier.

To evaluate the results of the classifiers while prioritizing effectiveness, we selected the **metrics** Macro-F1, Micro-F1, and Standard Deviation, as they provide a comprehensive assessment of model performance and ensure compatibility with the baseline. Macro-F1 ensures that all classes are considered equally important. Micro-F1 offers an overall view of performance, considering all instances. The Standard Deviation provides insights related to the consistency and robustness of the models, demonstrating the data distribution around the average of the Macro and Micro-F1 scores.

The execution of the experiments uses the following **methodology**. Based on the tabular subset used by the baseline, the heterogeneous graph was created using the specifications defined in step 1 of AGHE. In the subsequent steps 2 and 3, the proposed embedding compositions in this paper were generated. At the end of this process, we obtained the following embeddings: Metapaths, composition of Features and Metapaths, and composition of Aggregated Features and Metapaths. After creating the embedding compositions, we conducted four experiments to compare the metrics results with the baseline:

---

<sup>1</sup><https://www.kaggle.com/datasets/adrianmcmahon/imdb-india-movies>

1. Creation of the SVM model with Hold-Out and calculation of the scores for the embedding compositions;
2. Creation of the XGBoost and Ensemble models with Hold-Out and calculation of the scores for the embedding compositions;
3. Creation of the SVM, XGBoost, and Ensemble models with Cross-Validation and calculation of the scores for the embedding compositions;
4. Statistical analysis of the model that achieves the best result.

Our paper selected the SVM classifier because it applies to the data scenario we have in the experiments, which does not have a huge number of rows, a high capacity for generalization reducing the possibility of overfitting, and it is effective in high-dimensional spaces. The second reason is to maintain compatibility with the baseline, which also uses SVM. We introduced the XGBoost algorithm in our experiments because it performs very well on classification problems, also its implementation is based on decision trees, unlike the support vectors used by SVM. Additionally, our paper conducted experiments utilizing the Ensemble classification technique, incorporating multiple classifiers with diverse approaches. This combination aims to capture wide data patterns, thereby enhancing the robustness and predictive performance of the classification system. We performed calibration tests with various base classifiers and obtained the best results with the following ensemble combination: Random Forest Classifier, Extreme Gradient Boosting, Support Vector Machine, Decision Tree Classifier, Bootstrap Aggregating, Adaptive Boosting, Gradient Boosting, and Logistic Regression.

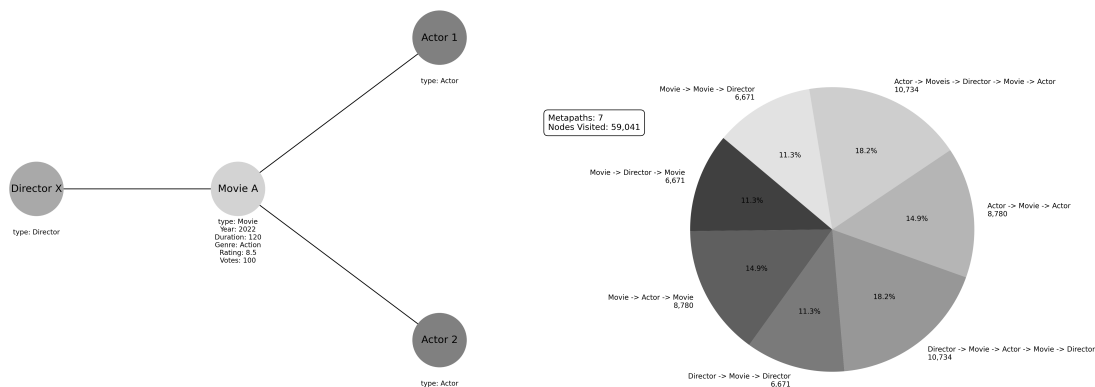
The baseline used many different percentages for training, but since the standard split for Hold-Out validation is 80% for training and 20% for testing, we decided to use it in our experiments. This decision ensures consistency and compatibility with the experiments performed in the baseline. In order to evaluate the results using a different approach to sample generation, our experiments were also guided by the Cross-Validation technique. We employed 5 iterations with 10 k-folds, applying stratified validation and shuffling the data before the division into folds. The metrics Macro-F1 and Micro-F1, which were also used in the baseline, were calculated in our paper by averaging the results and including their respective standard deviations.

## 6.2. Dataset Adaptation and Metapaths Definition

The heterogeneous graph schema after the transformation from IMDb movie table is shown in Fig. 3(a). Thus, the set of graph nodes, their fields, and the relationships provided by the edges correspond exactly to the fields from the original table. Genre is the field used as the multiclass of predictions, which is a little unbalanced where Drama at 45,3%, Action at 33,3%, and Comedy at 21,4%. Thus, we chose not to use the sampling approach, which involves adding or removing data. Instead, the calculated metrics for the classifiers used the “average” hyperparameter with the “weighted” value, which is also appropriate for unbalanced multiclass classification.

Metapaths are an important part of embedding compositions which were defined according to Equation 1 and are shown in Fig. 3(b). They are defined as follows:  $\mathcal{M} \leftarrow \{(Movie \rightarrow Director \rightarrow Movie), (Movie \rightarrow Actor \rightarrow Movie), (Director \rightarrow Movie \rightarrow Director), (Director \rightarrow Movie \rightarrow Actor \rightarrow Movie \rightarrow Director), (Actor \rightarrow Movie \rightarrow Actor), (Actor \rightarrow Movie \rightarrow Director \rightarrow Movie \rightarrow Actor), (Movie \rightarrow Movie \rightarrow Director)\}$ .





(a) Network schema data model of the heterogeneous graph created and used in this paper. (b) Distribution of graph nodes visited per metapath.

**Figura 3. Heterogeneous graph model and metapaths defined.**

### 6.3. Evaluation and Results

We present the evaluation results of our experiments, analyzing the performance metrics of the various models and comparing them to the baseline.

#### 6.3.1. Experiment 1

This experiment evaluates whether embedding compositions improve the results of the Node Classification task. We compare them with the baseline MAGNN using the same SVM classifier and Hold-Out training data. Table 1 shows the baseline achieved the best Micro-F1 performance with 61.53% using the Intra-Metapath embedding defined by itself. Using the same 80% training data, our paper slightly surpassed this, achieving 61.76% with the composition Features + Metapaths and 61.65% with the composition Aggregated + Metapaths.

**Tabela 1. MAGNN and SVM with Hold-Out and embedding compositions.**

Embeddings	MAGNN		SVM	
	Mac-F1	Mic-F1	Mac-F1	Mic-F1
Intra-Metapaths	61.44	<b>61.53</b>		
Metapaths			55.34	60.91
Features + Metapaths			55.84	<b>61.76</b>
Aggregated + Metapaths			55.82	<b>61.65</b>

#### 6.3.2. Experiment 2

For this experiment, we propose the use of XGBoost and Ensemble classifiers to evaluate whether embedding compositions enhance the results of the Node Classification task. We compare these results with the baseline using Hold-Out training. This experiment demonstrated that the XGBoost and Ensemble classifiers did not achieve better results

than the baseline or SVM models, according shown Table 2. Thus, the conclusion remains that SVM is the better algorithm to use for both single and embedding compositions.

**Tabela 2. MAGNN and classifiers with Hold-Out and embedding compositions.**

Embeddings	MAGNN		XGBoost		Ensemble	
	Mac-F1	Mic-F1	Mac-F1	Mic-F1	Mac-F1	Mic-F1
Intra-Metapaths	61.44	<b>61.53</b>				
Metapaths			54.19	58.16	53.93	<b>59.96</b>
Features + Metapaths			52.12	56.25	52.50	58.47
Aggregated + Metapaths			53.49	57.52	53.71	59.43

### 6.3.3. Experiment 3

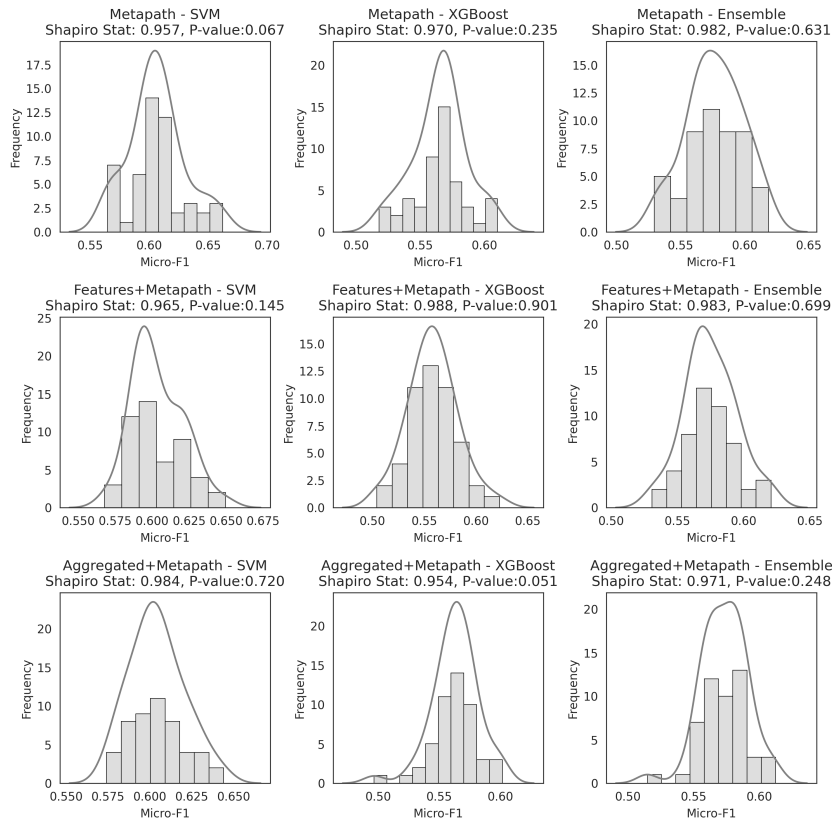
For this experiment, we propose creating SVM, XGBoost, and Ensemble models using Cross-Validation to evaluate whether embedding compositions improve the results of the Node Classification task when compared with baseline using SVM and Hold-Out. Table 3 shows the best result achieved, which was 65.89% Micro-F1, calculated from the embedding composition Aggregated + Metapaths, which is the highest result across all experiments, overcoming all models that used Hold-Out training data.

**Tabela 3. Classifiers with Cross-Validation and embedding compositions.**

Embeddings	SVM		XGBoost		Ensemble	
	Mac-F1	Mic-F1	Mac-F1	Mic-F1	Mac-F1	Mic-F1
Metapaths	60.16	64.19	58.67	62.29	56.56	61.65
Features + Metapaths	59.23	65.68	57.74	61.02	56.57	61.02
Aggregated + Metapaths	60.77	<b>65.89</b>	58.54	62.92	56.59	61.86

### 6.3.4. Experiment 4

It is the statistical analysis of the model that achieves the best result. Fig. 4 shows the data representation of the embedding compositions with the classifiers selected for the experiments, illustrating the data distribution and the statistical test results. We analyzed the Aggregated + Metapaths embedding composition with the SVM classifier, which achieved the best results in our experiments. However, a similar analysis applies to the other cases. We used Shapiro-Wilk Normality Statistical Test which is a reliable method for assessing whether data follows a normal distribution [Mohd Razali and Yap 2011]. This test is particularly sensitive to departures from normality and is suitable for small to medium sample sizes, making it ideal for our experimental conditions. The statistical tests achieved the following results: *Statistic W* - the value achieved of 0.984 is very close to 1, indicating that the data distribution fits well with the normal distribution; *P-value* - the value achieved of 0.720 is well above the common significance level, meaning there is not enough evidence to reject the null hypothesis (fail to reject H0) that the data follows a normal distribution.



**Figure 4. Statistical tests for data distribution.**

Therefore, we can conclude that the sample appears to be Gaussian, therefore, following a normal distribution, shown by the Kernel Density Estimation (KDE) curve in Fig. 4, confirmed the normality of the data, indicating the absence of outliers and that the mean is an appropriate measure of central tendency.

### 6.3.5. General Results Analysis

The results achieved through embedding compositions are encouraging, as they have proven to be effective, especially when used with the Cross-Validation technique, as demonstrated in Table 4, where all the models using Aggregated + Metapaths embedding composition outperformed the baseline used in this paper. Therefore, Table 4 presents only the best-performing Micro-F1 metric values along with their standard deviations, which highlight a near-linear distribution of the data around the means, concentrated between  $\pm 1\%$  and  $\pm 2\%$ . Consequently, demonstrating low variability, thus indicating that the model is reliable and robust to variations in the training data. Otherwise, we observe that all classifiers SVM, XGBoost, and Ensemble achieved their best performance metric when using the embedding composition Aggregated + Metapaths, achieving Micro-F1 of 65.89%, 62.92%, and 61.86%, respectively. The SVM model achieved the best performance using all embedding compositions, with 64.19%, 65.68%, and 65.89% for Metapaths, Features + Metapaths, and Aggregated + Metapaths. We can also evaluate the experiments from an ablation perspective, where the experiments aim to identify which embeddings and algorithms have the greatest impact on the effectiveness of the models.

**Tabela 4. The best Micro-F1 metric achieved with its standard deviation.**

Embeddings	SVM	XGBoost	Ensemble
	Mic-F1 STD	Mic-F1 STD	Mic-F1 STD
Metapaths	<b>64.19</b> 1.90	62.29 2.11	61.65 2.35
Features + Metapaths	<b>65.68</b> 1.74	61.02 1.99	61.02 2.24
Aggregated + Metapaths	<b>65.89</b> 2.03	<b>62.92</b> 2.12	<b>61.86</b> 2.13

The sequence of Macro-F1 60.77% and Micro-F1 65.89%, shown in Table 3, suggests that the model performs better overall than on average for each class. This can occur in scenarios with class imbalance, where the model is more effective in larger classes but struggles with smaller ones.

## 7. Conclusion

This paper explored the use of heterogeneous graphs and embedding compositions as key elements to enhance node representation. The experimental results based on the proposed embedding compositions were quite promising. By incorporating the proposed embeddings, particularly the Aggregated + Metapaths composition, our approach achieved outstanding results. The experimental outcomes demonstrated significant improvements in node representation and classification task. Additionally, our experiments using the proposed Cross-Validation technique, as an alternative to Hold-Out, achieved significantly better performance in the observed metrics. Therefore, we can conclude that the proposed embedding compositions are effective in enhancing node representation in heterogeneous graphs, consequently improving the results in subsequent applications.

Future works include evaluating other ReSys tasks, such as Link Prediction and Node Clustering; exploring different applications on the same heterogeneous graph, such as Community Detection, Fraud and Anomaly Detection; proposing a new node embedding composition using edge features; and implementing hyperparameters optimization to achieve the best classifier hyperparameter values for optimal classifiers performance.

## References

- Angonese, S. F. and Galante, R. (2024). AGHE: Approach for Generating Enhanced Heterogeneous Embeddings from Heterogeneous Graphs. In *51<sup>o</sup> Seminário Integrado de Software e Hardware (SEMISH)*, Brasilia, DF, Brazil.
- Dong, Y., Chawla, N. V., and 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, KDD '17*, page 135–144, New York, NY, USA. Association for Computing Machinery.
- Fu, X., Zhang, J., Meng, Z., and King, I. (2020). MAGNN: Metapath Aggregated Graph Neural Network for Heterogeneous Graph Embedding. In *Proceedings of The Web Conference 2020, WWW '20*, page 2331–2341, New York, NY, USA.
- Hamilton, W. L., Ying, R., and Leskovec, J. (2017). Inductive Representation Learning on Large Graphs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17*, page 1025–1035, Red Hook, NY, USA. Curran Associates Inc.

- Jin, D., Huo, C., Liang, C., and Yang, L. (2021). Heterogeneous Graph Neural Network via Attribute Completion. In *Proceedings of the Web Conference 2021, WWW '21*, page 391–400, New York, NY, USA. Association for Computing Machinery.
- Mohd Razali, N. and Yap, B. (2011). Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *J. Stat. Model. Analytics*, 2.
- Santana, D. and Ribeiro, L. (2023). Approximate similarity joins over dense vector embeddings. In *Anais do XXXVIII Simpósio Brasileiro de Bancos de Dados*, pages 51–62, Porto Alegre, RS, Brasil. SBC.
- Sun, Y. and Han, J. (2012). *Mining Heterogeneous Information Networks: Principles and Methodologies*. Morgan amp; Claypool Publishers.
- Sun, Y., Han, J., Yan, X., Yu, P. S., and Wu, T. (2020). PathSim: Meta Path-Based Top-K Similarity Search in Heterogeneous Information Networks. *Proc. VLDB Endow.*, 4(11):992–1003.
- Wang, X., Bo, D., Shi, C., Fan, S., Ye, Y., and Yu, P. S. (2023). A Survey on Heterogeneous Graph Embedding: Methods, Techniques, Applications and Sources. *IEEE Transactions on Big Data*, 9(2):415–436.
- Wu, S., Sun, F., Zhang, W., Xie, X., and Cui, B. (2022). Graph Neural Networks in Recommender Systems: A Survey. *ACM Comput. Surv.*, 55(5).
- Ying, R., He, R., Chen, K., Eksombatchai, P., Hamilton, W. L., and Leskovec, J. (2018). Graph Convolutional Neural Networks for Web-Scale Recommender Systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery amp; Data Mining, KDD '18*, page 974–983, New York, NY, USA. Association for Computing Machinery.
- Zhang, C., Song, D., Huang, C., Swami, A., and Chawla, N. V. (2019). Heterogeneous Graph Neural Network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery amp; Data Mining, KDD '19*, page 793–803, New York, NY, USA. Association for Computing Machinery.