# Innovation and Entrepreneurship Ecosystem Association Driven by Graph Neural Network

**Weiwei Yuan**
School of Information Engineering, Guangzhou Institute of Technology, Guangzhou 510075, Guangdong, China
Corresponding author: Weiwei Yuan (e-mail: 15920517416@163.com)

**ABSTRACT** The innovation and entrepreneurship ecosystem consists of multiple stakeholders, including governments, universities, and enterprises, and involves complex interactions among technological cooperation, investment relations, and policy support. To improve the representation and prediction capability of multi-entity and multi-relation ecosystem models, this paper introduces a heterogeneous graph neural network framework covering graph construction, feature representation, and relation-embedding modeling. A heterogeneous graph is first constructed with enterprises, universities, and governments as node types and technological cooperation, investment, and policy support as edge types. Structural and behavioral features of different nodes are then extracted and uniformly encoded into multimodal vectors. A heterogeneous graph attention network is subsequently introduced to conduct weighted modeling of node types and relationship semantics. Finally, low-dimensional node representations are generated through graph embedding for ecosystem subgroup partitioning, core node identification, and potential relationship prediction. Experimental results show classification accuracies of 0.92, 0.91, and 0.95 for enterprise, university, and government nodes, respectively. The AUC values for link prediction in technological cooperation, investment, and policy support relations all exceed 0.92. Node centrality correlations and semantic separation results further verify the effectiveness of multimodal embedding in preserving network structure and distinguishing heterogeneous entities.

**INDEX TERMS** innovation ecosystem, heterogeneous graph neural network, multi-entity relation modeling, graph attention network, link prediction

## I. INTRODUCTION

THE deep integration of global science, technology, and industry has elevated regional innovation capability as a key measure of national competitiveness. The innovation and entrepreneurship ecosystem consists of diverse entities—including research institutions, universities, enterprises, and policymakers—that interact through technological cooperation, capital flow, and policy support, forming a complex, dynamic network. This interconnectedness is especially vital in manufacturing sectors [1,2]. Analyzing its structure, identifying key nodes, and uncovering resource flow pathways are crucial for improving resource allocation, detecting bottlenecks, and shaping effective innovation policies, thereby supporting high-quality regional development [3,4].

While some studies have used machine learning to enhance prediction in network analysis, many still rely on manually crafted graph statistics, which depend heavily on expert knowledge and struggle with evolving, complex systems [5,6]. Graph Neural Networks (GNNs) have thus emerged for complex network analysis, automatically learning node representations by propagating information between connected nodes, with applications in social networks and bioinformatics [7–9]. As a key branch, HGNNs are particularly promising for modeling ecosystems, as they can natively handle networks with multiple node and edge types. Focusing on multi-agent, multi-relation modeling in innovation ecosystems, this paper proposes an HGNN-based framework to address challenges in heterogeneous structure representation. Using HAN as the core model, it first defines entity types and relationship semantics to clarify node attributes [10]. A multimodal encoder then integrates structural, behavioral, and textual features to enhance node expressiveness. During representation learning, an attention mechanism identifies important relation paths for weighted edge learning [11,12]. Finally, unified node embeddings are obtained for tasks such as subgroup detection, hub identification, and collaboration path prediction.

This study aims to develop an HGNN-based analytical framework to enhance structural modeling and predictive accuracy in innovation ecosystem analysis, addressing current limitations in data fusion and semantic expression. Using multi-source regional innovation data, this paper constructs a heterogeneous graph and learn node embeddings to support key tasks such as network identification, structural analysis, and collaboration path prediction [13,14]. The proposed method integrates graph learning with ecosystem theory, establishing a unified representation that incorporates multiple agents, relations, and modalities. This scalable approach offers a transferable pathway for analyzing heterogeneous complex systems, with promising applications.

## II. INNOVATION ECOSYSTEM STRUCTURE MODELING STRATEGY

Figure 1 shows the core technical route of the innovation ecosystem structure modeling strategy, which revolves around heterogeneous graph network modeling and includes four key links: heterogeneous graph construction, multimodal node feature encoding, HAN modeling, and graph embedding expression and node clustering. Each module is closely linked, reflecting the complete process from structure definition to semantic extraction, from relationship modeling to potential pattern discovery.

### A. CONSTRUCTING HETEROGENEOUS GRAPH NETWORK

#### 1) Node Type Mapping and Entity Standardization

Based on key entities in the regional innovation ecosystem, this study defines three node types for the heterogeneous graph: enterprises, academic institutions, and government agencies [15]. Node data are sourced from business registries, university directories, and government platforms. Entity names are normalized and disambiguated using a matching model based on Jaccard similarity and Levenshtein distance, with a merge threshold of 0.85. Each unified entity receives a unique ID and structural attributes for feature encoding. To maintain focus, the analysis is limited to these three core entity types, excluding other supporting entities such as incubators.

#### 2) Multi-Type Relationship Construction and Edge Weight Assignment

In terms of edge structure construction, this paper defines three types of edges based on observable interaction behaviors between entities: technical cooperation edges, investment relationship edges, and policy support edges. The construction of each type of edge is based on specific factual data to ensure the effectiveness of the edge and the accuracy of structural expression.

Technical cooperation edges are determined by joint patents and co-authored papers. If any two entities have jointly applied for patents or jointly published academic papers in the past three years, an undirected cooperation edge is established between them. To enhance the semantic expression of edges, each edge records the frequency of cooperation and serves as the basis for edge weights.

Investment edges are used to express situations where there are equity or capital relationships between enterprises, between enterprises and universities or between governments. Based on the investment subject, object, amount and round information disclosed in the investment and financing database, a directed edge (from the investor to the investee) is constructed, and the investment time and amount are recorded. To handle the long-tail characteristics of the amount distribution, the edge weight is logarithmically normalized and expressed as Formula 1:

$$w_{ij}^{\text{inv}} = \frac{\log(1 + \text{amount}_{ij})}{\log(1 + \max(\text{amount}))} \tag{1}$$

In formula 1, $\text{amount}_{ij}$ represents the total investment of investor $i$ in $j$.

The policy support edge represents the government's incentive behavior such as providing financial support and policy preference to enterprises or universities, which is a directed edge (from the government to the beneficiary). The edge weight is calculated using the following weighted scoring method, as shown in Formula 2:

$$\begin{aligned} w_{ij}^{\text{pol}} &= 0.5 \cdot \text{normalized amount} \\ &+ 0.3 \cdot \text{normalized duration} \\ &+ 0.2 \cdot \text{policy type score} \end{aligned} \tag{2}$$

In formula 2, normalized amount represents the normalized grant amount, normalized duration represents the normalized support period, and policy type score is assigned according to the support type: direct subsidy = 1.0, policy privilege = 0.8, indirect support = 0.6.

#### 3) Graph Structure Definition and Heterogeneous Graph Normalization

After completing the construction of nodes and edges, this paper constructs the overall heterogeneous graph structure $G = (V, E, T_V, T_E)$, where $V$ is the node set, $E$ is the edge set, $T_V$ is the node type set (including enterprise, academic, and government), and $T_E$ is the edge type set (including collaboration, investment, policy). To support subsequent HGNN training, all relationship edges are stored separately by type and a mapping index is established to ensure that the model can distinguish different semantic paths when calculating adjacency features [16,17]. All types of edge weights are normalized to avoid a single relationship weight dominating the embedding representation.

### B. MULTIMODAL NODE FEATURE ENCODING

#### 1) Node Structure Feature Extraction and Normalization

After completing the construction of the heterogeneous graph structure, in order to achieve effective representation of nodes in the graph neural network, it is necessary to encode the multidimensional attributes of different types of nodes into numerical vectors in a unified format. Taking into account the significant heterogeneity of structural attributes of various entities in the innovation and entrepreneurship ecosystem, this paper extracts multimodal features from

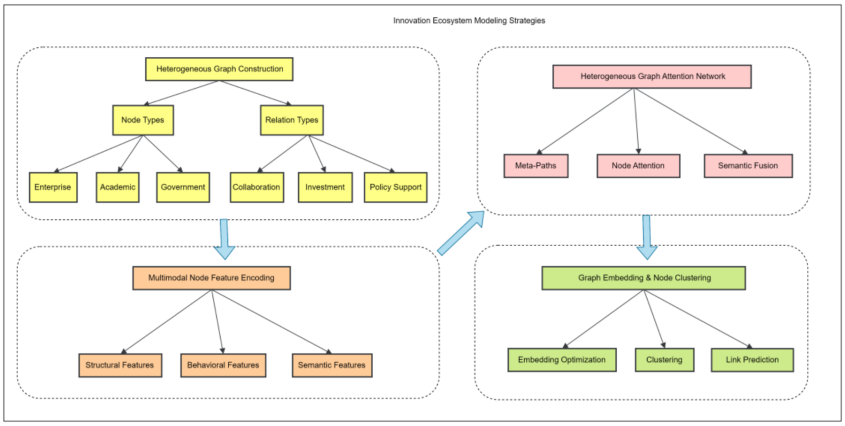


FIGURE 1. Innovation ecosystem structure modeling strategy framework

two levels: entity static attributes and interactive behavior attributes, and constructs a standardized input matrix for subsequent model training [18,19].

Among them, the text semantic features are extracted using a BERT pre trained model and reduced to 32 dimensions using PCA; All features (including large numerical features such as registered capital and small numerical features such as patent quantity) are subjected to logarithmic transformation and Z-score normalization (continuous variables) or Min-Max normalization (discrete variables) to ensure that the contributions of different dimensional features to the model are balanced and avoid a single feature dominating the learning process.

For enterprise nodes, structural attributes include registered capital (in 10,000 yuan), year of establishment, main industry classification, holding structure (whether there are universities/government holdings), and enterprise scale (divided into large, medium, and small and micro according to national statistical standards). Registered capital and year of establishment are numerical features, which are logarithmically transformed and then Z-score standardized. The main industry classification field is first coded according to GB/T 4754 for the first level industry, and then converted into a vector with a dimension of 20 after one-hot encoding. Holding structure and enterprise scale are discrete labels, which are embedded as dense vectors of length 4 and 3 respectively. The final structural vector dimension of the enterprise node is 32 dimensions.

The structural attributes of the government node are relatively simple, mainly including the institutional level (national/provincial/municipal), policy release frequency, annual fiscal allocation and the number of managed parks. The institutional level is represented by an embedded vector (dimension 4), the release frequency and the allocation are numerical variables, which are logarithmically transformed and Z-score normalized respectively. The number of managed parks is treated as a discrete value and is directly normalized.

To ensure the consistency of vector representation between different node types in subsequent models, all structural feature vectors are ultimately padded to the same dimension (set to 64 dimensions), and dimension alignment is accomplished using zero padding or linear projection.

Figure 2 visualizes the distribution and correlation of multimodal features. The left box plots show structural features are skewed with outliers, behavioral features are right-skewed with low fluctuation, and semantic features are concentrated with low variability, indicating clear distinctions beneficial for modeling. The correlation heatmap on the right reveals weak correlations between modalities, verifying that the multimodal design maintains the independence and representativeness of features, thus supporting its application in complex system modeling.

### 2) Behavioral Characteristics and Interaction Pattern Modeling

Node behavioral features in the heterogeneous graph are derived from observable interactions in multisource data. For enterprises, these include recent technical collaboration frequency, received investment events, and policy support. For universities, features cover joint research projects, technology transfer activity, and co-publication diversity. For governments, metrics include issued policy counts, coverage breadth, and scale. All indicators are normalized and combined into a 48-dimensional behavioral vector, which is then integrated with structural features to form the multimodal model input.

Beyond structural traits, node behavior patterns are also distinct and reveal their functional roles in cooperation. Therefore, this work extracts interactive behavior features to construct dynamic dimensions for each node type. For all nodes, this paper compute the frequency, temporal distribution, and intensity variation of their adjacent relationships from the heterogeneous graph, forming a multidimensional behavioral vector. Specifically, for each type of edge relationship (technical cooperation, investment, policy support), the cumulative number of edges (edge count), the average annual number of new edges (yearly increment), the edge weighted average (Edge weighted average), and the distribution entropy of the interacting party category (partner

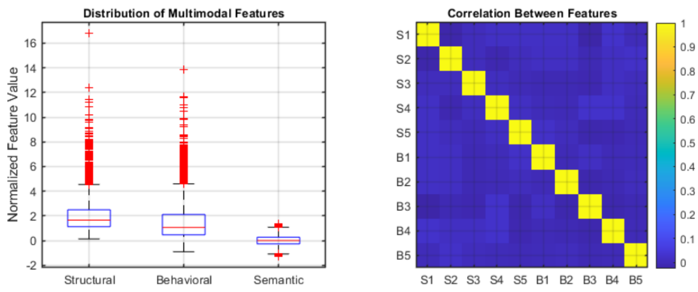


**FIGURE 2.** Visualization of multimodal feature distribution and correlation

entropy) are counted in the past three years. The distribution entropy is used to measure the diversity of the types of cooperation partners and is defined as Formula 3:

$$H_i = -\sum_{t \in T_V} p_i^t \cdot \log(p_i^t) \tag{3}$$

In formula 3, $p_i^t$ represents the interaction frequency ratio between node $i$ and node type $t$. This indicator is used to characterize the cross-type cooperation ability of nodes in behavior modeling. Each behavior indicator is standardized and concatenated to form a behavior feature vector.

For university and enterprise nodes, the clustering coefficient, node betweenness centrality and k-core value in the cooperation network are further introduced to quantify their structural functional potential in the ecological network [20,21]. Finally, the behavioral feature encoding dimension of all nodes is 48 dimensions, which are concatenated with the aforementioned structural features to form a complete multimodal node input vector. After a layer of nonlinear activation function (Rectified Linear Unit, ReLU), it is input into the graph neural network module. All feature encoding processes are completed before graph construction, generating a node feature matrix $X \in \mathbb{R}^{|V| \times d}$, which includes $d = 64$ (structure) $+48$ (behavior) $+32$ (semantics) $= 144$. This matrix serves as the input basis of the HGNN model to ensure the fusion expression capability of multi-source information.

### *C. MODELING WITH HAN*

This paper chooses Heterogeneous Graph Attention Network (HAN) instead of other non-hierarchical HGNN architectures, primarily because it can explicitly model higher-order associations under different semantic paths through meta-path mechanisms. In the innovation and entrepreneurship ecosystem, key relationships often span multiple node types. Such cross-type, multi-hop semantic paths cannot be effectively captured by ordinary GNNs or non-hierarchical heterogeneous models. HAN significantly improves its ability to model complex coupling relationships by assigning independent attention channels to each predefined meta-path and weighting and fusing the importance of paths at the semantic level.

#### 1) Meta-Path Construction and Semantic Channel Definition

After completing the encoding of heterogeneous graph structures and multimodal node features, this paper uses the Heterogeneous Graph Attention Network (HAN) to model multi-type nodes and their relationships. HAN is based on the meta-path mechanism and can extract structural semantic information from different types of relationship paths and achieve semantic discrimination through channel-level attention weights [22]. First, combining the above three types of edge types and three types of node types, five semantically significant meta-paths are constructed as the input channels of the HAN model, including: E-collab-E (enterprise-technical cooperation-enterprise), E-invest-E (enterprise-investment relationship-enterprise), G-policy-E (government-policy support-enterprise), A-collab-E (university-technical cooperation-enterprise), and G-policy-A (government-policy support-university).

#### 2) Local Attention Mechanism and Semantic Fusion

For each subgraph generated by a meta-path, a graph neural network encoding module based on the node attention mechanism is used to learn the local representation of the node under the semantic channel. The node representation initial vector is the multimodal vector $x_i \in \mathbb{R}^{144}$ constructed above. Each meta-path in HAN corresponds to a separate graph attention layer, which introduces a learnable attention coefficient to capture the importance of neighbor nodes to the central node. The attention weight $\alpha_{ij}$ between node $i$ and its neighbor $j$ is calculated as shown in Formula 4:

$$\alpha_{ij} = \frac{\exp\left(\mathrm{LeakyReLU}\left(a^T[Wh_i \| Wh_j]\right)\right)}{\sum_{k \in N_i} \exp\left(\mathrm{LeakyReLU}\left(a^T[Wh_i \| Wh_k]\right)\right)} \tag{4}$$

In formula 4, $W \in \mathbb{R}^{d' \times d}$ is the linear transformation matrix, $a$ is the attention weight vector, $\|$ represents vector concatenation, and $N_i$ is the neighbor set of node $i$. LeakyReLU is selected as the activation function, and the negative slope is set to 0.2. This mechanism allows the model to dynamically assign the weight of each neighbor in the information aggregation process and enhance the ability to distinguish structures. In order to stabilize training and improve expression ability, $K = 4$ parallel attention heads are used in each meta-path subgraph, each head independently calculates the attention distribution, and the output results are averaged and summarized as the semantic representation of the node. This design can improve the model's ability to express local structural diversity and reduce gradient oscillation during training. After obtaining the node representation $\{h_i^{(1)}, h_i^{(2)}, \ldots, h_i^{(M)}\}$ under each meta-path, the model is fused through the channel attention mechanism at the semantic level. In order to capture the importance difference between different paths, the semantic attention weight $\beta_m$ is introduced and calculated by the following Formula 5:

$$\beta_m = \frac{\exp\left(q^T \cdot \tanh(W_s \cdot \mu_m + b_s)\right)}{\sum_{l=1}^{M} \exp\left(q^T \cdot \tanh(W_s \cdot \mu_l + b_s)\right)} \tag{5}$$

$\mu_m = \frac{1}{|V|}\sum_{i=1}^{|V|} h_i^{(m)}$ represents the average representation of all nodes under the $m$th path, $W_s \in \mathbb{R}^{d' \times d'}$, $b_s \in \mathbb{R}^{d'}$, and $q \in \mathbb{R}^{d'}$ are semantic attention parameters, all of which participate in training optimization. The final node representation is the weighted sum of the outputs of each path, as shown in Formula 6:

$$h_i^{\text{final}} = \sum_{m=1}^{M} \beta_m \cdot h_i^{(m)} \tag{6}$$

This mechanism can effectively enhance the model's ability to distinguish semantic channels, avoid the introduction of interference information by irrelevant paths, and improve the efficiency of cross-type information integration.

#### 3) Model Parameter Setting and Regularization Optimization Strategy

The output dimension of each layer in the HAN model is set to d=64, and the entire model contains two layers of attention aggregation structure. ReLU activation function and dropout regularization term are used between each layer, and the dropout ratio is 0.5. The Adam optimizer is used for model training, the learning rate is set to 0.005, the upper limit of training rounds is 200 rounds, and the early stopping strategy is set to terminate when the validation set index has not improved for 15 consecutive rounds. To alleviate the risk of overfitting, the model introduces the L2 regularization term.

During the training process, the mini-batch sampling method is used to construct the subgraph, and the batch size is set to 1024 nodes. Considering the imbalance of node categories, the category weighted cross entropy is introduced into the loss function, and the category weight is adaptively calculated based on the node type and the subgroup distribution frequency to improve the recognition accuracy of the model in the minority subgroup.

The left figure of Figure 3 shows the distribution of attention coefficients of edge weight distribution in HAN. The overall distribution is biased to the left, indicating that the attention of most edges is relatively low and concentrated between 0.1 and 0.4. Only a few edges obtain higher weights during the model learning process, reflecting the ability of HAN model to distinguish key paths. The right figure of Figure 3 depicts the changing trend of loss value and verification accuracy during model training. It can be observed that the loss value drops rapidly in the first 50 rounds and stabilizes in the next 150 rounds of iterations, while the verification accuracy rises to nearly 1.

### D. GRAPH EMBEDDING EXPRESSION AND NODE CLUSTERING

#### 1) Node Embedding Generation and Representation Optimization

After completing the construction of the heterogeneous graph structure, node feature fusion and HAN modeling, this paper uses the final node representation $h_i^{\text{final}} \in \mathbb{R}^d$ calculated by the semantic fusion mechanism as the low-dimensional vector expression of the node in the embedding space for subsequent clustering, central node identification and link prediction tasks. The node embedding dimension $d$ is uniformly set to 128 and is learned end-to-end by the HAN model during training. To ensure the discriminability of the representation in spatial distribution, the model introduces an auxiliary clustering optimization term in the training objective and adds a structural consistency regularization term in addition to the node classification loss function, as defined in Formula 7:

$$L_{\text{embed}} = L_{\text{cls}} + \lambda \cdot L_{\text{intra}} \tag{7}$$

In Formula 7, $L_{\text{cls}}$ is the node label supervision loss (used for supervised training of partially labeled nodes), $L_{\text{intra}} = \sum_{i,j \in C_k} \|h_i - h_j\|^2$ represents the minimization of the distance between nodes of the same type, and $\lambda$ is the balance coefficient, which is set to 0.1. $C_k$ represents a set of nodes that share the same supervised functional category label, rather than predefined "enterprise/university/government" node types. This constraint aims to utilize the known functional labels of some nodes to guide the formation of embedding space and enhance the cohesion of similar nodes. This mechanism guides the node embedding vectors to naturally aggregate in the category space and improves the embedding quality. In order to avoid excessive shrinkage in the embedding space and ensure the separation between different node categories, this paper introduces the maximum inter-class divergence regularization term in the training process, and uses the minimum Euclidean distance between center points as a penalty indicator to enhance the structural contrast of node embedding. The final node embedding matrix $H \in \mathbb{R}^{N \times 128}$, where $N$ is the number of all nodes in the graph, is used as a unified representation for subsequent clustering and prediction modules to call. All embedding processes are accelerated by GPU, and the training convergence time is about 40 minutes. The final embedded vector is standardized to ensure stable distribution and is suitable for unsupervised task scenarios. Figure 4 shows the node embedding type and cluster visualization. The left figure of Figure 4 shows the

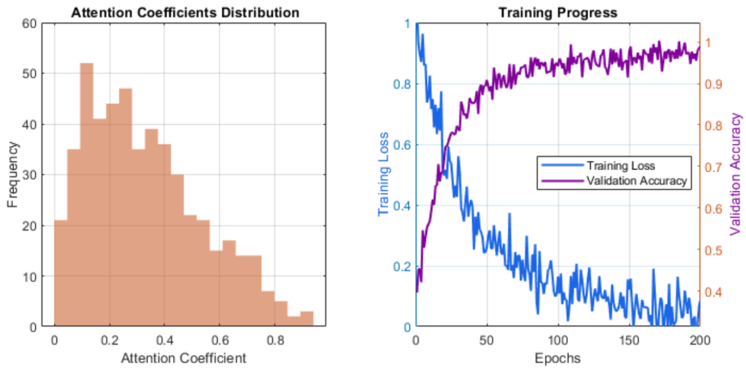


**FIGURE 3.** Attention distribution and training evolution of HAN model

distribution of node embeddings in two-dimensional space obtained by HGNN learning. Enterprise, university and government nodes are clustered in different areas, indicating that the embedding representation can effectively capture the characteristic differences of different types of subjects in the multi-relational ecological structure. The right figure of Figure 4 divides all node embeddings based on K-means clustering ($K = 5$), and it can be observed that the clustering boundaries of nodes in low-dimensional space are relatively clear. Some clustering results span across subject types, indicating that there is a potential coupling relationship across organizational types in the innovation ecosystem, which verifies the effectiveness of embedding vectors in discovering implicit structures and collaborative patterns within the system.

### 2) Subgroup Division and Core Node Identification

After obtaining the embedding vector H, this paper uses an unsupervised clustering method to identify structural subgroups in the ecosystem. Mini-Batch K-Means is selected as the clustering method to improve the clustering efficiency under large-scale graph embedding. The initial number of clusters K is determined based on the dual optimization of the Silhouette Coefficient and the Calinski-Harabasz index. Finally, a suitable K value is selected, and the clustering input is the normalized embedding vector. To enhance the clarity of clustering boundaries, the embedding space was reduced to 50 dimensions by PCA (Principal Component Analysis) before clustering to reduce the interference of redundant dimensions on the clustering structure. After clustering, the co-occurrence matrix and mutual information index were constructed to conduct a secondary test on the interaction edge density between subgroups, remove weakly connected nodes with fuzzy boundaries, and improve the structural stability of ecological subgroup division. In each subgroup, the node composition ratio and relationship density are further analyzed to identify highly cohesive subgraphs with functional independence and structural autonomy for subsequent structural evolution simulation and group behavior prediction. Based on the identification of subgroups, this paper introduces structural centrality indicators and information flow path indicators to jointly identify key hub nodes in the ecosystem. The node importance scoring function is defined as Formula 8:

$$S(i) = \alpha \cdot C_{\text{deg}}(i) + \beta \cdot C_{\text{bet}}(i) + \gamma \cdot \text{Comm}(i) \tag{8}$$

In formula 8, $C_{\text{deg}}(i)$ is the degree centrality of the node, $C_{\text{bet}}(i)$ is the betweenness centrality, and $\text{Comm}(i)$ represents the clustering consistency score of the node within the group. The weight parameters are $\alpha = 0.4$, $\beta = 0.4$, $\gamma = 0.2$, which are determined by cross-validation. This scoring method can identify core nodes with multi-channel connection capabilities and information forwarding functions in the ecosystem as the priority objects for subsequent policy guidance and capital layout.

### 3) Potential Relationship Prediction Mechanism

Based on the graph embedding results, this paper further completes the potential relationship prediction task, that is, predicting the node pairs that are not visible in the current graph structure but have a high probability of establishing edge relationships. The prediction task uses the edge pair similarity scoring method to calculate the potential connection probability $\hat{g}_{ij}$ of the node pair $(i, j)$ in the embedding space. The scoring function is defined as Formula 9:

$$\hat{g}_{ij} = \sigma\left(h_i^T W_p h_j + b\right) \tag{9}$$

In formula 9, $\sigma$ is the Sigmoid activation function, $W_p \in \mathbb{R}^{128\times128}$ is the training parameter matrix, and $b$ is the bias term. The training samples consist of actual edge relationships and randomly sampled unconnected node pairs in the graph, and the negative sampling ratio is set to 1:3.

In order to avoid edge type confusion, the prediction task independently trains three binary classification models on three types of edge relationships (technical cooperation, investment relationship, and policy support), and finally integrates and outputs a unified connection probability matrix for subsequent multi-objective tasks. The model training adopts the AUC optimization objective, and the loss function is binary cross entropy. The edge type embedding is introduced in the training process to enhance the model's ability to distinguish edge semantics.

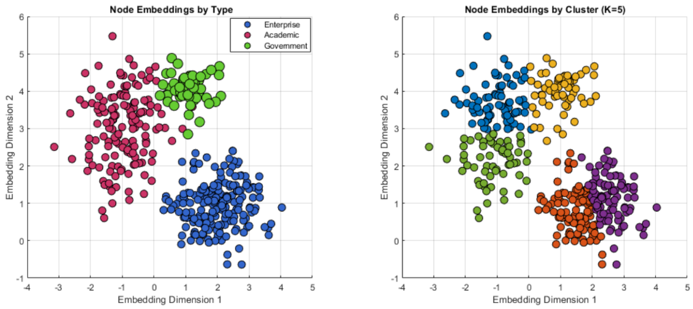


FIGURE 4. Node embedding type and cluster visualization

After the prediction module is trained, the potential edge probability can be generated for all unconnected node pairs, and the high-confidence connection relationship can be determined by setting a threshold ($\hat{g}_{ij} > 0.85$), which assists in completing the structure completion of the ecological network and identifying potential cooperation paths.

## III. ECOLOGICAL STRUCTURE IDENTIFICATION AND PREDICTION INDICATORS

This study constructs a heterogeneous graph based on real data from the Regional Innovation and Entrepreneurship Ecosystem Comprehensive Information Platform from 2018 to 2023. The graph contains 29,471 nodes (21,038 enterprises, 4,132 universities, 301 government agencies, and 4,000 other supporting entities) and 112,759 edges. In the construction and statistical description of the overall ecological network, the "other supporting entities" nodes are retained to reflect the overall ecological picture. But in the subsequent tasks of heterogeneous graph modeling, feature encoding, and core analysis, it has been excluded from node analysis as an independent type. The subsequent calculations in Table 2 aim to evaluate the semantic discrimination between core entity embeddings and such background nodes.

Edge types cover three main categories: technological cooperation, investment relationships, and policy support. Downstream tasks are evaluated using standard metrics: node classification task reporting accuracy, link prediction task area under the curve (AUC), subgroup partitioning quality is measured by modularity, and structure preservation is verified using Spearman centrality correlation coefficients. All results are based on five-fold cross-validation to ensure statistical robustness. The node classification task divided the dataset according to the establishment time of the institutions; the link prediction task used temporal partitioning to ensure that the timestamps of training edges are earlier than those of validation and test edges, and edges with direct adjacency to nodes in the training set were removed from the test set to prevent "neighbor leakage." The evaluation based on this strict partitioning effectively ensured the generalization reliability of the results.

The data used in this study comes from a de-identified research version of a provincial-level "Regional Innovation and Entrepreneurship Ecosystem Comprehensive Information Platform." This platform integrates multi-source data, including enterprise business registration information, university research projects, government documents, and investment and financing events. Data collection followed the platform's full-sample principle, covering all relevant entities and interaction records in the region from 2018 to 2023. Sensitive information has been anonymized using intervalization or hashing.

### *A. NODE CLASSIFICATION ACCURACY*

Node classification aims to predict functional subcategories within each node type (e.g., enterprise role / academic function / government agency level), rather than predefined node types. Node classification accuracy evaluates how well a model distinguishes nodes of known types, reflecting the separability of learned node embeddings. During training, some labeled nodes are held out as supervision. A classifier then predicts node categories, and accuracy is measured by alignment between predicted and true labels. To ensure robustness, the experiment uses five-fold cross-validation across enterprise, university, and government nodes, averaging results over random splits. This metric directly shows how well embeddings capture categorical boundaries, serving as a fundamental evaluation criterion. The evaluated models include HAN, GAT(Graph Attention Network), GraphSAGE(Graph Sample and Aggregation), GCN(Graph Convolutional Network), node2vec(Node to Vector), DeepWalk, LINE(Large-scale Information Network Embedding), SDNE(Structural Deep Network Embedding), Struc2Vec(Structural Similarity to Vector), HIN2Vec(Heterogeneous Information Network to Vector), Metapath2vec.

Figure 5 shows the classification performance of various graph neural network methods on different node types. The X-axis of the left subplot of Figure 5 represents the method name, including 11 mainstream models; the X-axis of the right subplot of Figure 5 represents the year and the Y-axis is the node classification accuracy. The left figure compares the accuracy of three types of nodes, namely enterprises, universities and governments, under different methods in the form of a bar graph. Overall, HAN performs best among all node types, with an accuracy of 0.92 for enterprise nodes, 0.95 for government nodes and 0.91 for university nodes. The accuracy of the three types of nodes is significantly better than the traditional random walk-based DeepWalk, node2vec and LINE methods. The right figure further shows the performance evolution of the HAN method on different nodes in the past six years, showing a stable and continuous upward trend, confirming the effectiveness of the heterogeneous attention mechanism in understanding multi-type nodes.

### B. MODULARITY SCORE

The modularity score evaluates the quality of community detection in a graph, gauging how well-separated and internally connected the subgroups are. After obtaining node groups through unsupervised clustering of embeddings, modularity is computed based on the actual edge connections in the original network. A high modularity indicates dense connections within groups and sparse links between them. In this study, modularity was calculated following Mini-Batch K-Means clustering, with the cluster number optimized per node category to balance structural distinction and stability. Different clustering methods include HAN Clustering, K-Means (K-means clustering), Spectral Clustering, DBSCAN (Density-Based Spatial Clustering of Applications with Noise), and GMM (Gaussian Mixture Model).

Figure 6 shows the performance of different clustering methods and node types under modularity analysis. In the left figure, the x-axis is the number of clusters K (ranging from 5 to 20), and the y-axis is the corresponding modularity score, which reflects the quality of community division of the clustering structure. It can be seen that the HAN clustering method has the best overall performance, and the modularity increases from 0.42 to 0.81 as the K value increases, which is better than traditional methods such as K-Means and spectral clustering. The right figure is a modularity heat map of different node types under $K = 7$ to 13. The x-axis is the number of clusters, the y-axis is the node type, and the value represents the modularity. The government node reaches the highest value of 0.77 when $K = 11$, indicating that its clustering structure is clearer and more stable.

### C. LINK PREDICTION AUC

Link prediction AUC (Area Under the Curve) measures a model's ability to identify potential connections between unlinked nodes. To evaluate this, some real edges are held out as positive test samples, while an equal number of non-existent node pairs are randomly sampled as negative examples. Similarity scores between node pairs are then computed from their learned representations to estimate connection likelihood. A higher AUC indicates better performance in distinguishing true from false edges. The evaluation was conducted separately for each relationship type—cooperation, investment, and policy—to thoroughly assess the model's semantic adaptability.

Figure 7 shows the link prediction performance of 11 models on three types of relationships. The horizontal axis is the name of the different models, and the vertical axis is the AUC score, which is used to measure the accuracy of the prediction. HAN leads in all three types of relationships. For example, its AUC for investment relationships is 0.94, and it also reaches 0.92 and 0.93 for technical cooperation and policy support, which are significantly higher than traditional methods such as DeepWalk (the lowest is 0.72) and LINE (the highest is only 0.78). Overall, heterogeneous perception models (such as HAN and Metapath2vec) show stronger expressive power and prediction accuracy when dealing with multi-type relationships. This verifies the effectiveness of modeling the attention mechanism of heterogeneous graphs, and shows that models that deeply integrate structural semantics have stronger link prediction generalization capabilities in complex innovation networks.

### D. NODE CENTRALITY CORRELATION

This metric assesses how well node embeddings preserve the structural importance of nodes from the original graph. First, structural centrality measures (e.g., betweenness, degree) are computed in the original graph. Corresponding centrality scores are then derived in the embedding space based on node distances or aggregation weights. Finally, the ranking consistency between the two sets of scores is evaluated. High correlation confirms that the embeddings retain nodes' structural roles, aiding in the identification of key hubs and validating the structural sensitivity of the representation for higher-level analysis.

Table 1 presents the centrality metrics (betweenness, degree, closeness) for three node types, comparing original graph values with those in the embedding space and assessing their correlation via Spearman's coefficient. Results show enterprise betweenness centrality is highly correlated (0.71, $p < 0.001$), confirming that embeddings preserve nodes' bridging roles. Academic and government nodes also show medium-to-high correlations (e.g., 0.62 for degree, 0.52 for closeness). These correlations demonstrate that the graph embeddings effectively capture structural features, validating the accuracy of the multimodal heterogeneous graph representation for understanding node roles in the innovation ecosystem.

### E. SEMANTIC SEPARATION

Semantic separation measures the clarity of boundaries between different node types in the embedding space, reflecting the model's ability to capture node heterogeneity. It is evaluated by computing the average vector distance between categories and the proportion of overlapping nodes across them. A higher separation score indicates that the model successfully distinguishes essential semantic differences and reduces representation mixing. This metric is especially suitable for heterogeneous graph tasks and serves as a key criterion for assessing whether multi-type semantic boundaries are well constructed.

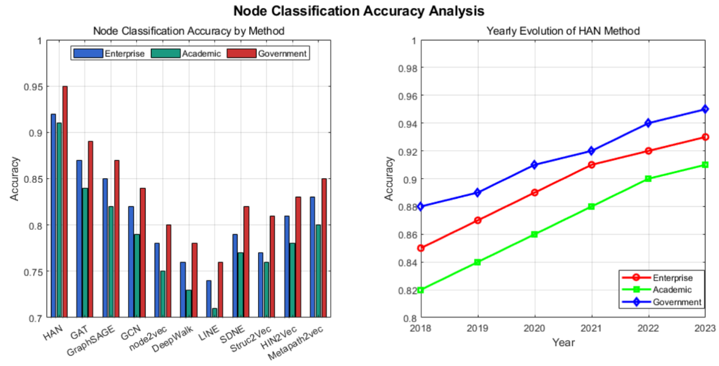


FIGURE 5. Classification performance of various graph neural network methods on different node types

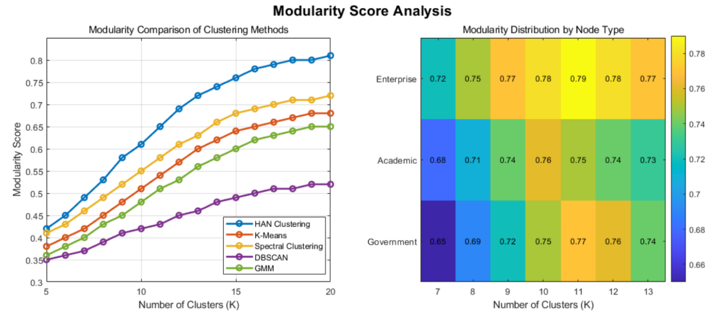


FIGURE 6. Performance of different clustering methods and node types

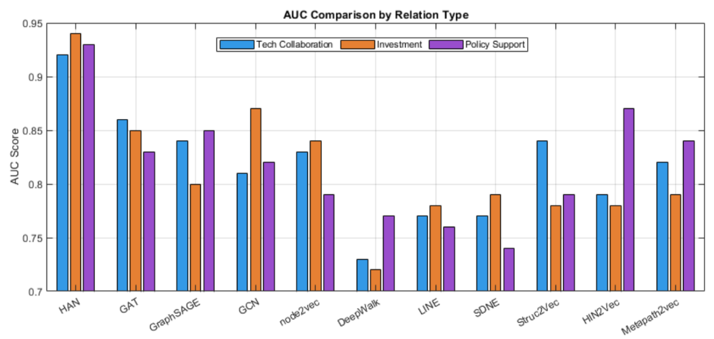


FIGURE 7. Link prediction performance of different models on three types of relationships

TABLE 1. Three centrality indicators of three types of nodes in the original graph and embedded space

| Node Type | Centrality Metric | Original Graph Centrality Value (Mean $\pm$ SD) | Embedding Space Centrality Score (Mean $\pm$ SD) | Spearman Correlation Coefficient | p-value | Note |
|---|---|---|---|---|---|---|
| Enterprise | Betweenness Centrality | $0.72 \pm 0.15$ | $0.68 \pm 0.13$ | 0.71 | $< 0.001$ | High Correlation |
| Enterprise | Degree Centrality | $0.65 \pm 0.18$ | $0.61 \pm 0.16$ | 0.68 | $< 0.001$ | High Correlation |
| Enterprise | Closeness Centrality | $0.58 \pm 0.12$ | $0.55 \pm 0.10$ | 0.63 | $< 0.001$ | Medium-High Correlation |
| Academic | Betweenness Centrality | $0.51 \pm 0.10$ | $0.48 \pm 0.09$ | 0.65 | $< 0.001$ | High Correlation |
| Academic | Degree Centrality | $0.42 \pm 0.08$ | $0.39 \pm 0.07$ | 0.62 | $< 0.001$ | Medium-High Correlation |
| Academic | Closeness Centrality | $0.47 \pm 0.11$ | $0.44 \pm 0.10$ | 0.6 | $< 0.001$ | Medium Correlation |
| Government | Betweenness Centrality | $0.38 \pm 0.07$ | $0.35 \pm 0.06$ | 0.58 | $< 0.001$ | Medium Correlation |
| Government | Degree Centrality | $0.33 \pm 0.05$ | $0.30 \pm 0.04$ | 0.55 | $< 0.001$ | Medium Correlation |
| Government | Closeness Centrality | $0.29 \pm 0.06$ | $0.27 \pm 0.05$ | 0.52 | $< 0.001$ | Medium Correlation |

TABLE 2. Semantic separation between pairs of nodes of different categories

| Category Pair | Average Vector Distance (Euclidean Distance) | Overlapping Node Ratio (%) | Semantic Separation Score (0–1) | Notes |
|---|---|---|---|---|
| Enterprise vs Academic | 4.32 | 8.7 | 0.85 | High Separation |
| Enterprise vs Government | 5.12 | 5.2 | 0.92 | Very High Separation |
| Academic vs Government | 3.87 | 10.1 | 0.81 | High Separation |
| Enterprise vs Other Support | 3.65 | 12.3 | 0.77 | Medium-High Separation |
| Academic vs Other Support | 3.41 | 14.5 | 0.72 | Medium-High Separation |

Table 2 presents the semantic separation between node category pairs using three metrics: average Euclidean distance, overlapping node ratio, and a semantic separation score. For instance, enterprises and governments show a clear distinction (distance: 5.12, separation: 0.92, overlap: 5.2%). In contrast, universities and other supporting entities exhibit closer proximity (distance: 3.41, separation: 0.72, overlap: 14.5%), indicating semantic overlap. Overall, the embedding space effectively captures both inherent differences and latent correlations among node types, aiding the interpretation of multimodal feature fusion.

## IV. CONCLUSIONS

This article constructs an innovation and entrepreneurship ecosystem analysis framework based on HGNNs, achieving systematic modeling of multiple subjects and relationships. This framework is particularly suitable for complex industrie, and can effectively reveal the laws of internal innovation collaboration and resource flow. By introducing attention mechanisms to fuse multi-source features, the model can accurately divide ecological subgroups, identify core nodes, and predict potential collaborations. The experiment verified the effectiveness of the model in maintaining structure and distinguishing semantics. The research results provide quantitative basis for the formulation of regional innovation policies and demonstrate practical potential in optimizing the supply chain of the industry.

## AUTHOR CONTRIBUTIONS

Weiwei Yuan designed, collected and analyzed the data, and drafted the manuscript. Weiwei Yuan conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Weiwei Yuan participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

## CONFLICTS OF INTEREST

The author declares no conflict of interest.

## FUNDING

This research received no external funding.

## ACKNOWLEDGEMENTS

Not applicable.

## REFERENCES

[1] R. C. F. Júnior, C. P. Lima, R. B. Viani, and A. S. da Fonseca, "The relationship between innovation ecosystem and bioeconomy: A bibliometric review," International Journal of Innovation, vol. 13, no. 2, pp. 1-26, 2025, doi: 10.5585/2025.26936.

[2] T. P. Haidegger, P. Galambos, J. K. Tar, M. Kozlovszky, Z. Zrubka, G. Eigner, et al., "Strategies and Outcomes of Building a Successful University Research and Innovation Ecosystem," Acta Polytechnica Hungarica, vol. 21, no. 10, pp. 13-35, 2024, doi: 10.12700/APH.21.10.2024.10.2.

[3] E. N. Yakovleva and N. A. Kamenskih, "Strategic Model of the University's Innovation Ecosystem," Vestnik Gosudarstvennogogumanitarno-Tekhnologicheskogo Universiteta, vol. (2), pp. 71-77, 2021.

[4] A. R. Silvestro, G. P. de Matos, D. J. C. da Silva, D. L. dos Santos Reis, and C. S. Teixeira, "An analysis on the interactions between actors of the innovation ecosystem and incubators," International Journal of Scientific Management and Tourism, vol. 10, no. 3, Art. no. e949-e949, 2024.

[5] T. Roundy P and A. Asllani, "Understanding AI innovation contexts: A review and content analysis of artificial intelligence and entrepreneurial ecosystems research," Industrial Management & Data Systems, vol. 124, no. 7, pp. 2333-2363, 2024, doi: 10.1108/IMDS-08-2023-0551.

[6] G. Giuggioli and M. M. Pellegrini, "Artificial intelligence as an enabler for entrepreneurs: A systematic literature review and an agenda for future research," International Journal of Entrepreneurial Behavior & Research, vol. 29, no. 4, pp. 816-837, 2023, doi: 10.1108/IJEBR-05-2021-0426.

[7] D. Wang, L. Han, L. Cong, H. Zhu, and Y. Liu, "Practical evaluation of human–computer interaction and artificial intelligence deep learning algorithm in innovation and entrepreneurship teaching evaluation," International Journal of Human–Computer Interaction, vol. 41, no. 3, pp. 1742-1750, 2025, doi: 10.1080/10447318.2023.2199632.

[8] N. Upadhyay, S. Upadhyay, and Y. K. Dwivedi, "Theorizing artificial intelligence acceptance and digital entrepreneurship model," International Journal of Entrepreneurial Behavior & Research, vol. 28, no. 5, pp. 1138-1166, 2022, doi: 10.1108/IJEBR-01-2021-0052.

[9] R. Widayanti and L. Meria, "Business modeling innovation using artificial intelligence technology," International Transactions on Education Technology (ITEE), vol. 1, no. 2, pp. 95-104, 2023, doi: 10.33050/itee.v1i2.270.

[10] D. Hao, "Hierarchical Analysis and Graph Attention Network for University Students' Innovation and Entrepreneurship Learning Platform," Informatica, vol. 48, no. 19, pp. 145-158, 2024, doi: 10.31449/inf.v48i19.6545.

[11] S. Wang, "The relationship between economic growth and carbon emissions based on the combination of graph neural network and wavelet transform," Energy Informatics, vol. 8, no. 1, pp. 1-23, 2025, doi: 10.1186/s42162-025-00521-9.

[12] D. Hain, R. Jurowetzki, S. Lee, and Y. Zhou, "Machine learning and artificial intelligence for science, technology, innovation mapping and forecasting: Review, synthesis, and applications," Scientometrics, vol. 128, no. 3, pp. 1465-1472, 2023, doi: 10.1007/s11192-022-04628-8.

[13] L. Jain, R. Katarya, and S. Sachdeva, "Opinion leaders for information diffusion using graph neural network in online social networks," ACM Transactions on the Web, vol. 17, no. 2, pp. 1-37, 2023, doi: 10.1145/3580516.

[14] A. Benghalem, A. S. Bennihi, and M. Benbouziane, "Artificial Intelligence and Entrepreneurship: A Comprehensive Bibliometric and Text Mining Analysis of Global Research Trends," Les Cahiers du Mecas, vol. 21, no. 1, pp. 1-17, 2025, [Online]. Available: https://asjp.cerist.dz/en/article/268132.

[15] C. Troise, C. Ben-Hafaïedh, M. Tani, and S. A. Yablonsky, "Guest editorial: New technologies and entrepreneurship: Exploring entrepreneurial behavior in the digital transformation era," International Journal of Entrepreneurial Behavior & Research, vol. 28, no. 5, pp. 1129-1137, 2022, doi: 10.1108/IJEBR-08-2022-999.

[16] Z. Huang and G. Liu, "Prediction model of college students entrepreneurship ability based on artificial intelligence and fuzzy logic model," Journal of Intelligent & Fuzzy Systems, vol. 40, no. 2, pp. 2541-2552, 2021, doi: 10.3233/JIFS-189247.

[17] F. Bertoni, S. Bonini, V. Capizzi, M. G. Colombo, and S. Manigart, "Digitization in the market for entrepreneurial finance: Innovative business models and new financing channels," Entrepreneurship Theory and Practice, vol. 46, no. 5, pp. 1120-1135, 2022, doi: 10.1177/10422587211038480.

[18] B. B. Gupta, A. Gaurav, and P. K. Panigrahi, "Analysis of the development of sustainable entrepreneurship practices through knowledge and smart innovative based education system," International Entrepreneurship and Management Journal, vol. 19, no. 2, pp. 923-940, 2023, doi: 10.1007/s11365-023-00853-6.

[19] D. K. Kanbach, L. Heiduk, G. Blueher, M. Schreiter, and A. Lahmann, "The GenAI is out of the bottle: generative artificial intelligence from a business model innovation perspective," Review of Managerial Science, vol. 18, no. 4, pp. 1189-1220, 2024, doi: 10.1007/s11846-023-00696-z.

[20] L. Grilli and M. Pedota, "Creativity and artificial intelligence: A multilevel perspective," Creativity and innovation management, vol. 33, no. 2, pp. 234-247, 2024, doi: 10.1111/caim.12580.

[21] Z. Jumayeva, "Scientific and methodological foundations of utilizing motivational factors in developing entrepreneurial activities," International Journal of Artificial Intelligence, vol. 1, no. 4, pp. 1571-1574, 2025.
[22] Y. Yang, "One-class graph moderating attention neural network in quality assessment of creative ideas," Neural Computing and Applications, vol. 36, no. 7, pp. 3369-3388, 2024, doi: 10.1007/s00521-023-09256-8.