Journal of Computer and Communications, 2026, 14(9), 51-70
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Identifying Key Nodes in Urban Public Transportation Networks Using Multi-Feature Fusion and GraphSAGE Network: A Case Study in Lanzhou City-China

Mingxia Qu, Huifang Feng*

College of Mathematics and Statistics, Northwest Normal University, Lanzhou, China
Email: *hffeng@nwnu.edu.cn

How to cite this paper: Qu, M.X. and Feng, H.F. (2026) Identifying Key Nodes in Urban Public Transportation Networks Using Multi-Feature Fusion and GraphSAGE Network: A Case Study in Lanzhou City-China. *Journal of Computer and Communications*, **14**, 51-70.
https://doi.org/10.4236/jcc.2026.149005

Received: August 12, 2026
Accepted: September 20, 2026
Published: September 23, 2026

Abstract

Identifying key nodes in urban public transportation networks is crucial for optimizing transportation flow and enhancing the resilience of transportation networks. Existing methods often suffer from shortcomings such as insufficient integration of global and local structural information and limited accuracy in ranking node importance. To address these issues, a multi-feature fusion model is proposed for identifying key nodes in urban public transportation networks based on the Graph Sampling and Aggregation Network (GraphSAGE). First, multi-dimensional node features are constructed from three dimensions including the network topology, operations, and spatial distribution. Second, local neighborhood features are captured by the GraphSAGE network, and a global context extraction module is constructed, through which global context information for nodes is extracted via weighted pooling layers and iterative expansion. Furthermore, a multi-layer neural network is employed to predict node importance, a gravity model is used to fuse multiple centrality metrics to generate reference labels. Finally, a ranking loss function is adopted to train the network model. An empirical analysis is conducted on the Lanzhou transportation network. The results show that the proposed method achieves a Kendall's τ of 0.8568, representing a 0.86% improvement over the best baseline method, thereby validating its superiority in ranking accuracy. Vulnerability analysis further confirms the effectiveness of the proposed method in identifying key nodes. The results can provide an effective basis for decision-makers to formulate risk-resilient contingency plans for urban public transportation networks.

Keywords

Urban Public Transportation Networks, Key Node Identification, GraphSAGE, Multi-Dimensional Features, Gravity Model

1. Introduction

In real life, many systems can be modeled as complex networks, such as transportation systems, power systems, and social networks [1]-[3]. In these networks, accurately identifying key nodes is crucial for understanding network structure, optimizing network performance, and improving system resilience. Urban public transportation networks, the core infrastructure for city operations, become increasingly complex with rapid urbanization and growing transportation demand, making key node identification especially essential for optimizing transportation flow and enhancing network resilience when resources are limited [4].

Numerous methods have been proposed to identify key nodes, including degree centrality, betweenness centrality, k-shell decomposition, and eigenvector centrality [5]. Although these methods provide valuable insights into network structure, they all suffer from inherent limitations, and accurate identification is often difficult to achieve when relying solely on a single metric. The TOPSIS multi-attribute decision-making method [6] was used to integrate multiple centrality metrics and generate a comprehensive importance ranking. Moreover, the introduction of machine learning and deep learning techniques has further advanced key node identification. Sun *et al.* [7] proposed a model based on the fusion of a Transformer encoder and a CNN, which significantly improved the accuracy of node influence rankings by aggregating multi-hop neighborhood features. Xiong *et al.* [8] generated node representations via a graph convolutional autoencoder, effectively fusing local and global structural information, and combined these representations with list ranking loss optimization to enhance node importance prediction accuracy.

Complex network theory provides a framework for analyzing transportation networks [9]. He *et al.* [10] proposed an NK-shell method based on neighborhood k-shell values, which improves upon the traditional k-shell decomposition by integrating the k-shell values of first-order neighbors and the number of second-order neighbors. This method has been validated in a bus-subway composite network and outperforms existing methods in terms of ranking monotonicity and identification accuracy. Ye Qing [11] conducted a topological vulnerability assessment of Beijing's rail transit system; the results indicated that functional failures at hub stations could trigger network-wide cascading failures. Regarding the analysis of urban transportation network topologies, Chen Feng *et al.* [12] selected three types of indicators—node degree, node betweenness, and edge betweenness—to construct an evaluation system, and systematically assessed network reliability by analyzing changes in the shortest path length and the size of the largest connected component following node failures. Zhang *et al.* [13] proposed a framework for identifying key stations that combines hierarchical computation with cross-layer fusion, constructing a multi-time-period multi-layer network model to capture pattern heterogeneity and intraday structural dynamics.

To overcome these limitations, recent studies have incorporated actual traffic demand and data-driven methods to further improve identification accuracy. Wang *et al.* [14] further proposed a calculation method based on transportation demand, using passenger flow distribution results driven by actual origin-destination (OD) matrices to replace purely topological weights, thereby achieving more precise identification of key nodes in multimodal networks. Lin *et al.* [15] established a public transit network model based on complex network theory, comprehensively considering travel resistance and path selection probability to determine network weights, and adopted the ant colony algorithm to solve the optimization problem with improved network efficiency as the objective function. Zhao *et al.* [16] proposed a method for identifying key nodes in transportation networks based on a multi-channel graph convolutional network with gated attention. Through multi-channel feature extraction and adaptive fusion, this method achieved performance superior to traditional centrality methods and graph neural network (GNN) baselines on multiple real-world datasets. Yin *et al.* [17] proposed an improved mutual information (IMI) method for identifying key nodes in urban rail transit networks, accounting for cross-dependencies between stations, and effectively identified key nodes with low degree but strong dependencies in the Chengdu Metro case study.

Although there are already many methods for identifying key nodes, existing methods still face limitations. Traditional topology-based approaches rely on static network structures and struggle to integrate multi-source features such as operational and spatial characteristics, while deep learning methods tend to favor either a local or global perspective, lacking a unified framework. To address these issues, a key node identification model based on GraphSAGE and multi-feature fusion is proposed for urban public transportation network in this paper. Node features are constructed from three dimensions—topological features, operational characteristics, and spatial distribution. The model captures both local and global structural information through the collaboration of GraphSAGE and a global context extraction module, fuses multiple centrality metrics and generates node labels via a gravity model, and is optimized using a ranking loss function. Experiments on the Lanzhou transportation network demonstrate that the proposed method can effectively identify key nodes.

2. Transportation Network Construction

Urban composite public transportation networks consist of multiple modes such as buses, subways, light rail, and trams. In this paper, an urban public transportation network system composed of buses and subways is considered, and the Space-L method is adopted to construct the urban public transportation network topology. Bus stops and subway stations are abstracted as nodes, and the links between adjacent stations on the same line are abstracted as edges, forming an unweighted and undirected network. All geographic coordinates are recorded in the WGS-84 geographic coordinate system (EPSG: 4326). To ensure physically meaningful distance measurements, all station coordinates are projected onto the Universal Transverse Mercator (UTM) coordinate system, with each city assigned to its corresponding UTM zone, resulting in metric units (meters). As shown in Figure 1, the comprehensive urban public transportation system combining subways and buses is abstracted as a two-layer coupled network $G_{SB} = \{V, E, S\}$, where $S = \{S_S, S_B\}$ is the set of layers, with S_S denoting the subway layer and S_B denoting the bus layer. The node set is defined as $V = V_S \cup V_B$, where V_S is the set of subway stations and V_B is the set of bus stops. The edge set is defined as $E = E_L \cup E_C$, where $E_L = E_{SS} \cup E_{BB}$ is the set of intra-layer edges, and E_C is the set of inter-layer coupling edges.

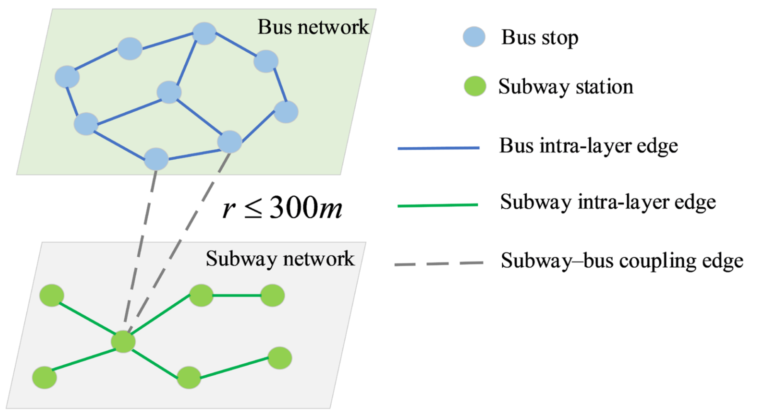

Figure 1. Construction of the coupled subway-bus network.

Within each layer, undirected edges are established based on the adjacent station relationships determined by the lines. If multiple edges exist between two nodes, they are merged into a single edge. To establish inter-layer connections, the Euclidean distance between station coordinates in the UTM plane is adopted as a conservative approximation of walking distance, and a spatial coupling criterion is defined based on the service radius requirements of rail transit feeder connections specified in GB/T 51328-2018 Standard for Urban Comprehensive Transport System Planning [18]. Specifically, an inter-layer coupling edge is established between a bus stop $v_b \in V_B$ and a subway station $v_s \in V_S$ if their Euclidean distance is less than the 300-meter transfer threshold, that is,
$E_C = \{(v_s, v_b) \mid d(v_s, v_b) < 300\ \text{m}, v_s \in V_S, v_b \in V_B\}$. Duplicate stops are eliminated at the source by directly using the stop uniqueness files provided by the China Public Transport Operation Network Dataset (CPTOND-2025) [19], in which each unique stop is assigned a distinct stop_id. Stops with similar geographic coordinates but different stop_id values are retained as separate nodes to avoid incorrect merging that could disrupt route topology and line affiliation information. If the con-

structed network contains multiple connected components, only the largest connected subgraph is retained for subsequent analysis.

3. Methodology

3.1. Node Multi-Feature Construction

To comprehensively characterize the structural properties of nodes in urban public transportation networks, multi-dimensional node features are constructed from three perspectives: network topology, operational characteristics, and spatial distribution.

1) Network topology features

K-core number, eccentricity, structural hole, and h-index are selected for capturing the network topology features. The k-core number [20] indicates a node's depth within the network backbone; eccentricity [21] measures its maximum distance to other nodes; structural hole [22] quantifies the potential to occupy an intermediary position; and the h-index [23], derived from degree distribution, supplements node influence assessment within its neighborhood.

2) Operational features

The operational features include the number of serving routes, route diversity, average operating duration, whether the station is a subway station, and whether the station is a transfer station. Route diversity is defined as the number of distinct routes serving a stop, counted by grouping the route_cn field in the CPTOND stops table by stop_id and removing duplicates, measured in counts. Average operating duration is the mean operating duration across all routes serving a stop, computed from the start_time and end_time fields in the routes table (in HHMM format), with cross-day handling applied when end_time is earlier than start_time, measured in hours. Transfer station status is a binary indicator. Metro stations are marked as 1, and bus stops within 300 m of a metro station (*i.e.*, having transfer edges) are marked as 1; otherwise 0. Transfer edges are identified by the 300 m Euclidean distance threshold in the UTM plane. The number of serving routes and route diversity jointly reflect a station's line connectivity; average operating duration indicates its service time span; and the binary indicators of subway or transfer station status provide modal attributes, facilitating the analysis of passenger capacity variations across modes and the structural significance of coupling nodes in the two-layer network.

3) Spatial distribution features

Spatial location, station coverage, nearest neighbor distance, average neighbor distance, and spatial dispersion of neighbors are considered for capturing the spatial distribution features. These features are derived from station coordinates using UTM projection, yielding units in meters. Specifically, station coverage is computed as a weighted sum of the number of stops within 500 m, 1000 m, and 2000 m radii (weights: 0.5, 0.3, 0.2), normalized by the total number of nodes, yielding a dimensionless value between 0 and 1. Spatial dispersion of neighbors is the mean pairwise distance among all neighboring stops, measured in kilometers; if the number of neighbors is less than 2, this value is set to 0.

The above features originate from three different data domains, with significant disparities in numerical scales and distribution characteristics. Therefore, in the preprocessing stage, Z-score standardization is performed separately for each city to eliminate the influence of dimensional differences on model training. For a node v_i its multi-dimensional feature vector is denoted as:

$$x_i^{multi} \in R^N \tag{1}$$

where N is the number of multi-dimensional features. The multi-dimensional feature matrix of all nodes in the transportation network is expressed as:

$$X^{multi} = \left[x_1^{multi}, x_2^{multi}, \cdots, x_n^{multi}\right] \in R^{n \times N} \tag{2}$$

where n is the total number of nodes in the network.

A learnable linear projection layer is applied to map the feature matrix X^{multi} to the target embedding dimension space. Layer normalization is employed to improve training efficiency, reduce sensitivity to initial weights, enhance generalization, and mitigate overfitting. The specific computation is formulated as:

$$X_{proj} = \mathrm{LN}\left(X^{multi} W_{proj} + b_{proj}\right) \tag{3}$$

where $\mathrm{LN}(\bullet)$ denotes the layer normalization function, $W_{proj} \in R^{N \times d}$ is the learnable weight matrix, $b_{proj} \in R^d$ is the bias term, and d is the embedding dimension. The projected feature $X_{proj} \in R^{n \times d}$ serves as the input feature for subsequent modules.

3.2. GraphSAGE-Based Feature Extraction

Graph Sample and Aggregate (GraphSAGE) [24] is an inductive graph representation learning framework that updates node representations by aggregating features from their neighbors, offering high computational efficiency and strong scalability. At layer l, for a node v_i, a fixed-size neighborhood set $N(i)$ of size S is randomly sampled from its neighbor set $N_S(i)$, and the neighborhood features are aggregated as,

$$h_{N_i}^{(l)} = \mathrm{AGGREGATE}\left(h_j^{(l-1)}\right), \forall j \in N_S(i) \tag{4}$$

Then, the node's own features are concatenated,

$$h_i^{(l)} = \sigma\left(W^{(l)}\left[h_{N_i}^{(l)} \,\|\, h_i^{(l-1)}\right] + b^{(l)}\right) \tag{5}$$

where σ denotes the ReLU activation function, and $W^{(l)} \in R^{d \times 2d}$ is the weight matrix of the l-th layer. The aggregator function $\mathrm{AGGREGATE}(\bullet)$ is implemented as the mean aggregator in this paper,

$$\mathrm{AGGREGATE}\left(h_j^{(l-1)}\right) = \frac{1}{S} \sum_{j \in N_S(i)} h_j^{(l-1)}, \forall j \in N_S(i) \tag{6}$$

A three-layer GraphSAGE network is adopted to extract structural features of nodes, followed by layer normalization to improve training stability. The overall computation can be formulated as:

$$H_{SAGE} = \sigma\Big(\mathrm{GraphSAGE}\Big(\sigma\Big(\mathrm{GraphSAGE}\Big(\sigma\big(\mathrm{GraphSAGE}\big(X_{proj}\big)\big)\Big)\Big)\Big)\Big) \tag{7}$$

where σ is the ReLU activation function, and $H_{SAGE} \in R^{n \times d}$ is the output of the GraphSAGE branch.

Through multi-layer neighbor aggregation, this pathway progressively disseminates node features to farther neighborhoods, thereby effectively capturing the global structural information of nodes.

3.3. Global Context Extraction

To capture the global structural information of the entire transportation network and compensate for the information limitations that may arise from local aggregation in GNNs, a global context extraction module is designed based on learnable scoring and weighted pooling. This module takes the output features of the GraphSAGE network $H_{SAGE} \in R^{n \times d}$ as input, computes a contextual importance score for each node through a two-layer MLP with Tanh activation, and then obtains weighting coefficients via Softmax normalization. Finally, a global context vector is derived through weighted pooling and expanded to each node, enabling all nodes to share the same global perspective.

First, the fused node features are fed into a two-layer MLP with Tanh activation to compute an importance score $e \in R^{n \times 1}$ for each node,

$$e = \sigma_1\left(\sigma_1\left(H_{SAGE}W_1 + b_1\right)W_2 + b_2\right) \tag{8}$$

where σ_1 denotes the Tanh activation function $W_1 \in R^{d \times d/2}$ and $W_2 \in R^{d/2 \times 1}$ are weight matrices, and $b_1 \in R^{d/2}$ and $b_2 \in R$ are bias terms.

Then, the score vector e is normalized by Softmax to obtain the normalized weights $w \in R^{n \times 1}$,

$$w = \mathrm{softmax}(e) \tag{9}$$

Based on these weights, weighted pooling is applied to the node features to obtain the global context vector $c \in R^d$,

$$c = \sum_{i=1}^{n} w_i \cdot h_{SAGE}^{(i)} \tag{10}$$

where $h_{SAGE}^{(i)} \in R^d$ denotes the fused feature vector of the i-th node.

Finally, the context vector is broadcast to all nodes by repeating it n times, yielding the context matrix $C \in R^{n \times d}$ as the output of this module.

$$C = 1_n \cdot c \tag{11}$$

This module allows each node's final representation to capture both local neighborhood information and global graph context, thereby providing a more comprehensive characterization of node roles within the network and richer features for node importance prediction.

3.4. Key Node Identification

The features $H_{SAGE} \in R^{n \times d}$ output by the GraphSAGE network module and the features $C \in R^{n \times d}$ output by the global context extraction module are concatenated along the feature dimension to obtain the final feature matrix,

$$H_{final} = \left[H_{SAGE} \,\|\, C\right] \in R^{n \times 2d} \tag{12}$$

Subsequently, H_{final} is fed into a two-layer MLP to obtain the prediction score $\hat{y} \in R^{n}$,

$$\hat{y} = \sigma_2\left(\mathrm{MLP}\left(H_{final}\right)\right) \tag{13}$$

where σ_2 denotes the Sigmoid activation function, $\mathrm{MLP}(\bullet)$ denotes a fully connected network. A higher score $\hat{y}$ indicates a more important node within the network.

3.5. Node Label Generation

To generate ranking labels for node importance, a two- stage label generation method is adopted that integrates centrality metrics and the gravity model. This method first computes node importance by fusing multiple centrality metrics, and then characterizes the influence of nodes on their surrounding nodes via the gravity model. The final result serves as the importance label for nodes. The specific steps are as follows.

1) Multi-dimensional Centrality Fusion

To characterize node importance, five classic centrality metrics are selected, Degree Centrality (DC), Eigenvector Centrality (EC), PageRank (PR), Closeness Centrality (CC) [6], and NK-shell [10]. Among them, NK-shell extends the traditional k-shell by incorporating first-order and second-order neighbor information, more comprehensively reflecting a node's core hierarchy and local influence. To eliminate scale differences among these metrics, Min-Max normalization is applied to the above five metrics, and they are fused with equal weights to obtain the node mass for each node.

2) Gravity-based Importance Calculation

Node mass reflects the importance of individual nodes; however, to account for the mutual influence between nodes, the gravity model is further introduced [25], [26]. By treating nodes as mass points, the model evaluates the comprehensive influence of nodes within the network through simulating gravitational interactions between them.

The shortest path distance is adopted as the distance metric between nodes. The shortest path distances between all node pairs are computed to obtain the distance matrix $D = \left[d_{ij}\right]_{n \times n}$, where d_{ij} denotes the shortest path distance between nodes v_i and v_j. To reduce computational complexity and focus on local influence, the gravity radius is set to $R = 3$, and only node pairs with $d_{ij} \le R$ are retained. The gravity importance S_i of node v_i is defined as,

$$S_i = \sum_{j: d_{ij} \le R, j \ne i} \frac{M_i \cdot M_j}{d_{ij}^2} \tag{14}$$

where M_i and M_j denote the node masses of v_i and v_j, respectively. The gravity importance vector of all nodes is given by,

$$S = \left(S_1, S_2, \cdots, S_n\right)^{\mathrm{T}} \in R^n \tag{15}$$

3) Label Generation

Min-Max normalization is applied to the gravity importance scores S_i to obtain the final labels,

$$y = \frac{S - \min(S)}{\max(S) - \min(S) + \varepsilon} \tag{16}$$

where ε is a tiny constant to prevent division by zero. The i-th component $y_i \in [0,1]$ represents the importance label of node v_i.

3.6. Training Optimization

The loss function is denoted as $Loss$. In this paper, a pairwise ranking loss is adopted to enable the model to learn the relative importance ranking among nodes. This loss is consistent with the hinge loss formulation in Ranking SVM [27]. As node pairs are randomly sampled, the loss function is defined as,

$$Loss = \frac{1}{|P|} \sum_{(i,j) \in P} \max\left(0, \delta - \mathrm{sign}\left(y_i - y_j\right)\left(\hat{y}_i - \hat{y}_j\right)\right) \tag{17}$$

where P is the set of randomly sampled node pairs, y_i and y_j denote the ground-truth importance scores of nodes i and j, $\hat{y}_i$ and $\hat{y}_j$ denote the predicted scores of nodes i and j, $\delta > 0$ is a margin hyperparameter, and $\mathrm{sign}(\cdot)$ denotes the sign function. This loss function, by minimizing $Loss$, enables the model to accurately fit node importance while preserving the correctness of relative order between nodes, thereby improving the model's performance in ranking tasks.

3.7. Framework of the Proposed Model

The proposed model for identifying key nodes in urban public transportation networks is shown in **Figure 2**. The model comprises six core modules: transportation network construction, node multi-feature construction, GraphSAGE-based feature extraction, global context extraction, key node identification, and node label generation. First, features from multi-source data are constructed based on three dimensions: topological structure, operational status, and spatial distribution. Next, a GraphSAGE network is developed to capture local neighborhood features of the network. Furthermore, a global context extraction module is designed to extract global contextual information of nodes through weighted pooling and repeated expansion, and a multi-layer neural network is employed to predict node importance. Meanwhile, a gravity model is used to fuse multiple centrality metrics to generate reference labels, and a ranking loss function is adopted to train the network model. Finally, the trained network is applied to identify key nodes in the urban public transportation network.

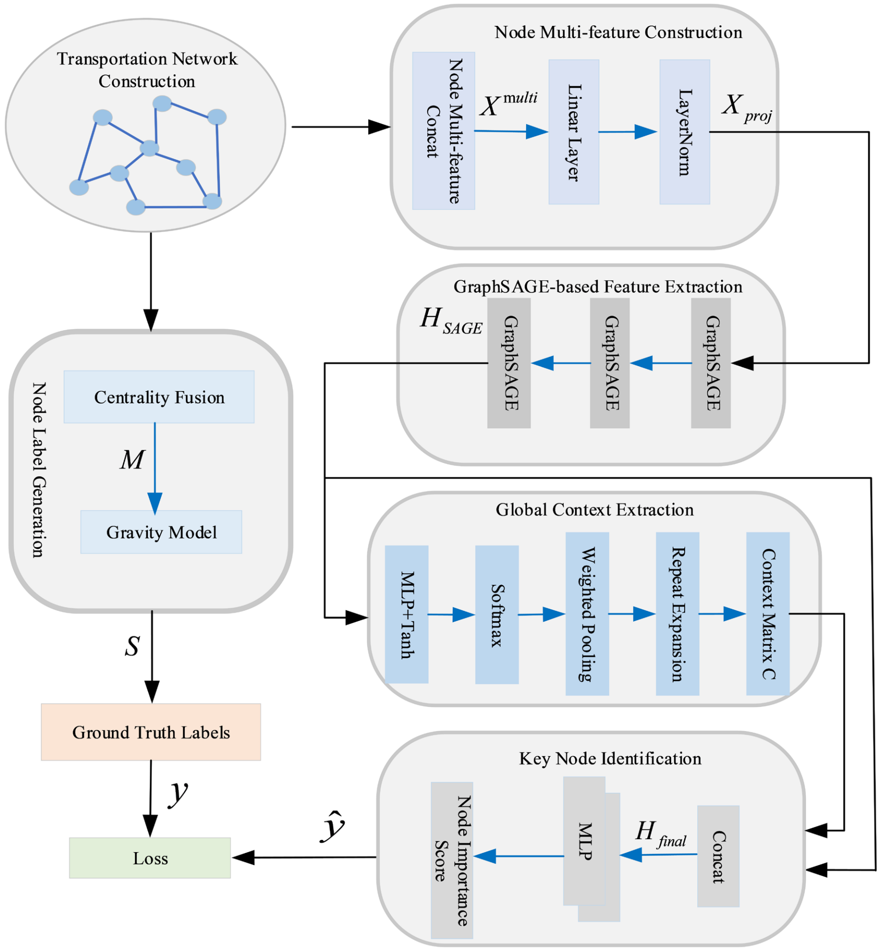

Figure 2. Framework of the proposed model.

4. Experiment Set

4.1. Datasets

The public transportation network data used in this paper is sourced from the China Public Transport Operation Network Dataset (CPTOND-2025) [19]. This dataset covers the subway networks of 351 cities in China and includes extensive information such as routes, stations, and operational attributes. To date, it is the first comprehensive nationwide public transportation dataset in China, covering 351 cities, some of which have both bus and subway networks, while others have only bus networks.

A total of 9 cities were selected to construct the experimental dataset. The training set comprised 7 cities (Xining, Yinchuan, Urumqi, Hohhot, Taiyuan, Beijing, and Chengdu). Wuhan was the validation set, and Lanzhou was the test set. This ensures that the training and test data are geographically independent, thereby effectively evaluating the model's generalization ability. In terms of network scale, taking Lanzhou as an example, the final network contains 1143 nodes (1116 bus stops and 27 metro stations) and 1884 edges (1794 bus edges, 27 metro edges, and 63 transfer edges), with an exclusion ratio of 2.47%.

4.2. Evaluation Metrics

1) Kendall's τ Correlation Coefficient

Kendall's τ correlation coefficient [28] is a statistical measure of ordinal association between two ranked lists, defined as,

$$\tau = \frac{c-d}{\frac{1}{2}n(n-1)} \tag{18}$$

where n denotes the number of nodes, c represents the number of concordant pairs, and d represents the number of discordant pairs.

In this study, the reported Kendall's τ measures the ordinal agreement between the predicted node rankings and the gravity-centrality surrogate labels, rather than an independent observation of real-world node importance. In the absence of ground-truth importance labels for urban public transportation networks, these surrogate labels—generated by fusing multiple centrality measures with a gravity model—serve as a reasonable proxy for node importance, providing a quantitative basis for comparing different ranking methods.

2) Largest Connected Component Ratio

The Largest Connected Component (LCC) [29] ratio is a metric for evaluating network vulnerability. This metric assesses the effectiveness of node importance ranking methods by measuring the change in the size of the largest connected component after the network undergoes attacks. The LCC ratio is defined as,

$$\mathrm{LCC}(\rho) = \frac{|C_{after}|}{|C_{initial}|} \tag{19}$$

where $|C_{initial}|$ is the number of nodes in the largest connected component of the original network, and $|C_{after}|$ is the number of nodes in the largest connected component after removing a fraction ρ of nodes. $\mathrm{LCC}(\rho) \in [0,1]$. This paper adopts the LCC metric to validate the Vulnerability of the key node rankings identified by the model. A faster decline in the LCC value indicates that the nodes identified by the model are more destructive to the network structure.

4.3. Baseline Models

To evaluate the effectiveness of the proposed model, Betweenness Centrality, Clustering Coefficient, TOPSIS, GCN, FGM, GCNGM, and SOLG are selected as baseline models.

Betweenness Centrality [6]. Betweenness centrality is a global centrality metric based on shortest paths, which quantifies the frequency with which a node appears on the shortest paths between other nodes, thereby measuring the node's bridging role in the network.

Clustering Coefficient [6]. The clustering coefficient measures the degree to which the neighbors of a node are also neighbors of each other, reflecting the local clustering and information transfer efficiency of the network.

TOPSIS [6]. TOPSIS is a multi-attribute decision-making method that ranks alternatives by calculating their relative closeness to the positive and negative ideal solutions. In node importance evaluation, nodes are treated as alternatives and centrality metrics as attributes, with a higher closeness indicating greater node importance.

GCN [30]. Graph Convolutional Network (GCN) performs feature extraction and node representation learning on graph-structured data through spectral graph convolution, enabling the simultaneous aggregation of node features and topological information.

FGM [31]. FGM is a node importance evaluation method based on the gravity model. It defines node mass by integrating multiple attributes such as degree, k-shell, and eigenvector centrality, and evaluates node importance in conjunction with shortest path distances.

GCNGM [32]. GCNGM is a gravity model-based node importance evaluation method that leverages node embedding representations. This method generates node topological embeddings via a fixed-weight graph convolutional network, uses the L2 norm of the embedding vectors as node mass, and constructs a gravity model combined with shortest path distances to evaluate node importance.

SOLG [33]. SOLG is a node importance evaluation method that integrates multi-stage topological features. It quantifies the connection strength between nodes by extending the Jaccard coefficient, and fuses first-order, higher-order, and global influences to assess node importance.

4.4. Experimental Setup

The number of GraphSAGE layers is set to 3, and all subsequent experiments are conducted under this configuration. The hidden layer dimension in GraphSAGE is set to 128. The margin hyperparameter δ is set to 0.1. The AdamW optimizer is employed with an initial learning rate of 5×10^{-4}, and a weight decay of 1×10^{-4}, The total number of training epochs is set to 500. For model selection, we adopted early stopping on the Wuhan validation set, and the best model parameters were saved for final testing. Lanzhou data were used exclusively as the test set and were never involved in model training, validation, or hyperparameter tuning.

For all baseline models, the hyperparameters suggested by the original authors are adopted, further adjusted based on the dataset, and all methods are ensured to be compared under identical training conditions. All experiments are conducted on a rented cloud server equipped with an NVIDIA RTX 3090 (24 GB) GPU, 14 vCPUs Intel Xeon Gold 6330 CPU, 90 GB RAM, and 80 GB SSD storage, running Ubuntu 20.04. The experiments are implemented using PyTorch 1.11.0, Python 3.8, and CUDA 11.3.

5. Experimental Results and Analysis

5.1. Ranking Performance Comparison

Figure 3 shows a comparison of the node importance ranking performance between the model proposed in this paper and the baseline models on the Lanzhou public transportation network. Overall, there are significant differences in ranking accuracy among the various types of methods. First, traditional centrality metrics perform poorly in the ranking task, reflecting that relying solely on the static topological structure or local features of the network makes it difficult to accurately identify key nodes in the public transportation network. Among the multi-feature fusion methods and supervised learning baseline models, the Kendall's τ results for TOPSIS, FGM, and GCN fall in the middle range. Among these, the GCN model achieved a Kendall's τ of 0.6703, indicating the basic suitability of graph neural networks for extracting spatial association features of nodes in urban public transportation networks. Among the unsupervised learning baseline methods, the SOLG algorithm and the GCNGM algorithm, which is based on graph convolutional and gravity models, performed relatively well, with Kendall's τ of 0.8495 and 0.8213, respectively, significantly outperforming the previous methods. In the Lanzhou public transportation network test, the proposed model achieved the highest Kendall's τ of 0.8568. Compared with the optimal baseline model, SOLG, represents a relative improvement of approximately 0.86%. These results indicate that the proposed model demonstrates better ranking accuracy in the task of identifying the importance of transportation nodes and can effectively identify key nodes within the network.

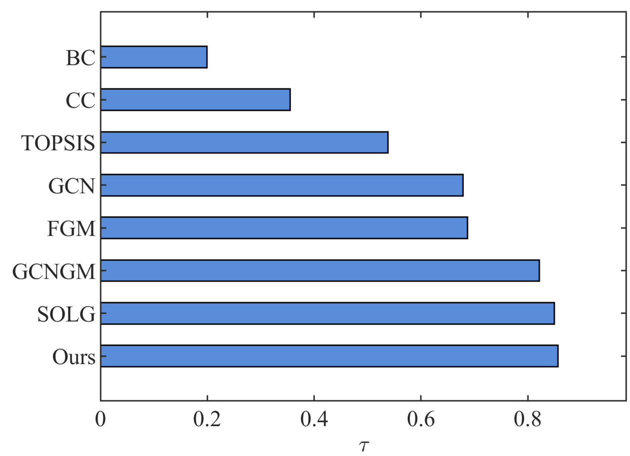

Figure 3. Comparison of Kendall's τ among different methods.

5.2. Correlation Analysis

To analyze the consistency between the predicted scores of each method and the ground-truth labels, Figure 4 presents the scatter plots of prediction scores versus label values for all methods on the Lanzhou dataset. In these plots, the horizontal axis denotes the node importance prediction scores produced by each method, and the vertical axis denotes the computed label values.

An examination of the scatter distributions reveals that the points of the proposed model are the most tightly clustered around the diagonal, indicating a strong linear correlation between the predicted scores and the labels. For the baseline methods SOLG and GCNGM, although the overall distribution generally aligns with the diagonal, underestimation is observed in certain regions, causing the corresponding points to fall slightly below the diagonal. The scatters of GCN and FGM also exhibit a discernible diagonal trend, but notable deviations can be observed in some areas. In stark contrast, the results of BC, CC, and TOPSIS display a scattered and stacked pattern, failing to effectively discriminate node importance. In summary, the scatter pattern of the proposed model on the Lanzhou transportation network demonstrates superior consistency compared with the baseline methods, thus validating its effectiveness in the node importance assessment task.

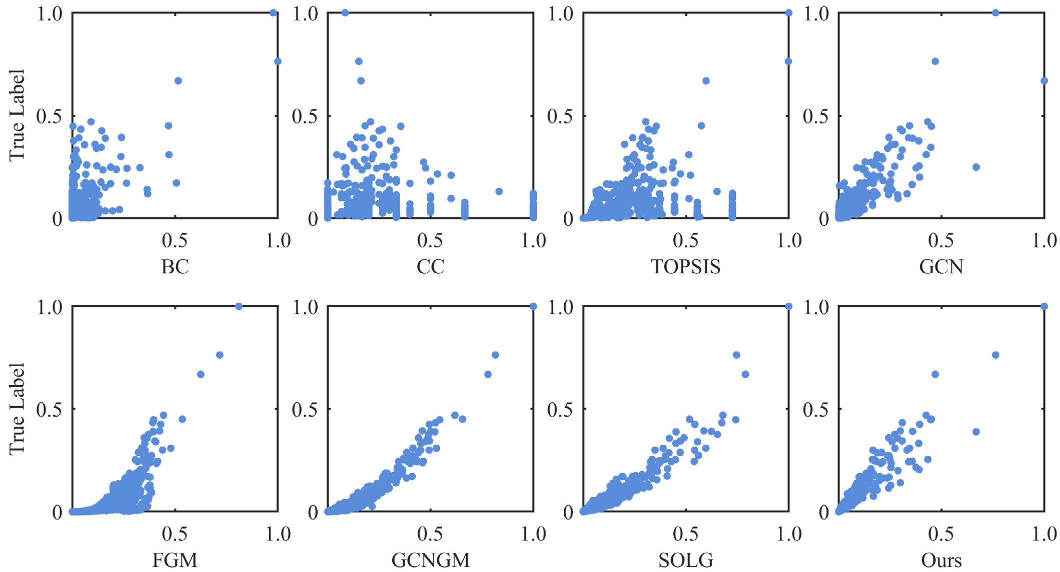

Figure 4. Scatter plots of predicted scores versus ground-truth labels for all methods.

5.3. Key Node Identification

To determine node importance levels, the K-means clustering algorithm (K = 12) is applied to the importance scores produced by the proposed model. Based on the average score of each cluster, the 12 clusters are further grouped into three importance levels: the cluster with the highest scores (30 nodes) is labeled as key nodes, the cluster with the second-highest scores (34 nodes) as important nodes, and the remaining 3rd to 12th clusters collectively as ordinary nodes (1079 nodes), totaling 1143 nodes. Figure 5 illustrates the spatial distribution of node importance levels in the Lanzhou public transportation network. Spatially, key nodes exhibit an east-west belt-like distribution and are highly concentrated in the core commercial areas of the main urban district, whereas the peripheral route endpoints are predominantly ordinary nodes. This distribution pattern closely matches the constrained elongated urban form of Lanzhou, shaped by the Yellow River valley, thus verifying the practical effectiveness of the model.

Table 1 presents the specific ranking of the 30 key nodes identified by the model. Among them, "Public Transport Group," "Xizhan Shizi," and "Wenhua Palace" are located at intersections of core commercial areas and main arterial roads; "Lanzhou West Railway Station" and "Lanzhou Railway Station" serve as key transportation transfer hubs; and "Provincial Government" and "Lanzhou University" function as essential public service and activity centers. The accurate identification of these nodes as key further validates the applicability and reliability of the proposed method in real-world public transportation networks.

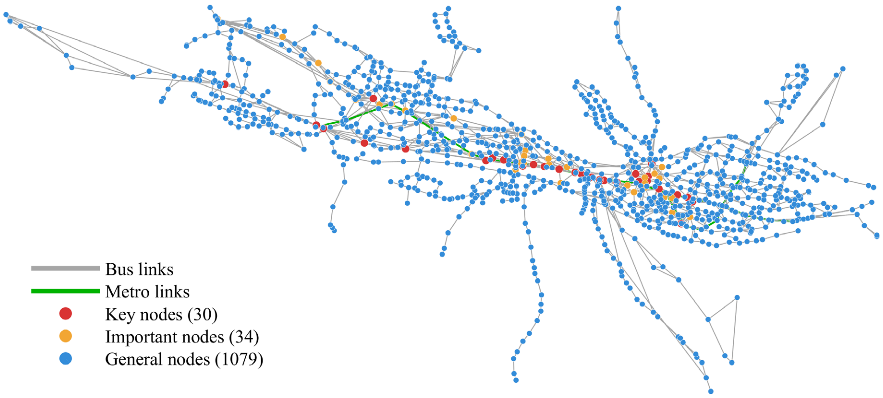

Figure 5. Spatial distribution of identified key nodes.

Table 1. Top 30 key nodes identified in Lanzhou.

Rank	Station name	Rank	Station name	Rank	Station name
1	Caochangjie	11	Qilihe Bridge	21	Shengou Bridge
2	Public Transport Group	12	Lanzhou West Railway Station East Entrance	22	Xihu Park
3	Wenhua Palace	13	Dongfanghong Square West Entrance	23	Jiefangmen
4	Xizhan Shizi	14	Xiguan Shizi	24	Lanzhou University
5	Xiaoxihu	15	Provincial Government	25	City Government
6	Qin'an Road Primary School	16	Lanzhou Railway Station	26	Provincial Academy of Agricultural Sciences
7	Lanzhou West Railway Station	17	Municipal Committee	27	Aoti East Street South Entrance
8	Dongfanghong Square East Entrance	18	Hongmenzi Bus Stop	28	Panxuan Road West Entrance
9	Liujiaying Crossroad	19	Xiuchuan New Village	29	Provincial Museum
10	Lanzhou West Railway Station North Square	20	Western Market	30	Chenguanying

5.4. Vulnerability Analysis

To verify the practical effectiveness of the proposed model in identifying key nodes, a Vulnerability analysis is conducted using a sequential node removal strategy and observing the decay of the relative size of the largest connected component (LCC). Figure 6 illustrates the curves of the LCC ratio under different node ranking methods on the Lanzhou public transportation network, where the horizontal axis represents the proportion of nodes removed in descending order of importance, and the vertical axis represents the relative size of the largest connected component.

As shown in Figure 6, in the initial node removal interval from 0% to 30%, the LCC ratio of the proposed model declines most steeply. This indicates that removing nodes sequentially according to the importance ranking identified by the proposed model can dismantle the network connectivity at the fastest rate. The unsupervised methods SOLG and GCNGM also exhibit rapid decay in the early stage, though their overall performance is slightly inferior to that of the proposed model. Notably, the decline of GCN is slightly slower than the above three methods, with a stagnation in the LCC curve observed between the removal ratios of 40% and 50%. This reflects the over-reliance of the GCN architecture on local information during feature aggregation, which hinders its ability to accurately identify key nodes connecting different regions. In contrast, the traditional single centrality metrics (BC, CC) and the multi-attribute aggregation framework TOPSIS present relatively gentle decline curves, suggesting that such methods struggle to comprehensively capture the nodes that play a critical role in the overall connectivity of transportation networks.

Overall, the Vulnerability performance of the proposed model is consistent with the conclusions drawn from the Kendall's τ analysis, further validating its effectiveness and reliability in the task of key node identification.

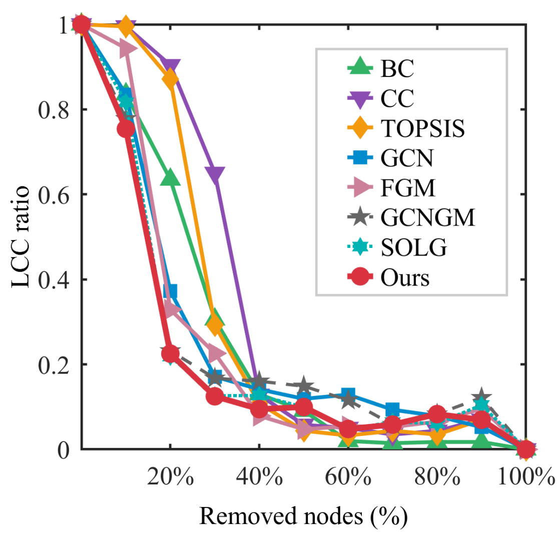

Figure 6. LCC ratio versus node removal proportion for different methods.

6. Conclusions

The problem of identifying key nodes in urban public transportation networks was addressed by proposing a model based on multi-feature fusion and the GraphSAGE network. The proposed model integrated multi-dimensional node features—including network topology, operational characteristics, and spatial distribution—to characterize nodes, employed a GraphSAGE network to extract local neighborhood features, combined multi-centrality and gravity models to generate reference labels, and adopted a pairwise ranking loss function for training optimization. An empirical analysis conducted on the Lanzhou public transportation network demonstrates the effectiveness of the proposed model. Experimental results showed that the method achieved an average Kendall's τ of 0.8568, representing a relative improvement of approximately 0.86% over the best baseline method, thereby demonstrating the model's significant advantage in ranking accuracy. Vulnerability analysis further validated the method's effectiveness. These research findings provided effective technical support for the evaluation of node importance in public transportation networks using deep learning and offered a solid basis for decision-makers to formulate network risk mitigation plans.

This study still has certain limitations. The current model primarily relies on static network topology and operational characteristics as its main inputs and has not yet incorporated dynamic transportation flow data, making it difficult to characterize the temporal patterns of node importance during peak hours and emergency scenarios. Furthermore, its applicability to larger-scale public transportation networks in terms of computational efficiency requires further validation. In future work, transportation flow data are planned to be incorporated to calibrate the model and enhance its interpretability, thereby providing more transparent decision-making support for transportation planning.

Acknowledgements

This research is supported in part by National Natural Science Foundation of China (NSFC) (Grant No. 72461030).

Author Contributions

Conceptualization, Mingxia Qu and Huifang Feng; Methodology, Mingxia Qu and Huifang Feng; Software, Mingxia Qu; Validation, Mingxia Qu and Huifang Feng; Formal analysis, Mingxia Qu; Investigation, Mingxia Qu; Resources, Mingxia Qu; Data curation, Mingxia Qu; Writing—original draft preparation, Mingxia Qu; Writing—review and editing, Mingxia Qu and Huifang Feng; Visualization, Mingxia Qu; Supervision, Huifang Feng; Project administration, Huifang Feng; Funding acquisition, Huifang Feng. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Szaciłło, L., Jacyna, M., Szczepański, E. and Izdebski, M. (2021) Risk Assessment for Rail Freight Transport Operations. *Eksploatacja i Niezawodność—Maintenance and Reliability*, **23**, 476-488. https://doi.org/10.17531/ein.2021.3.8

[2] Rozhkov, A. (2024) Applying Graph Theory to Find Key Leverage Points in the Transition toward Urban Renewable Energy Systems. *Applied Energy*, **361**, Article ID: 122854. https://doi.org/10.1016/j.apenergy.2024.122854

[3] Hua, Z., Jing, X. and Martínez, L. (2023) Consensus Reaching for Social Network Group Decision Making with ELICIT Information: A Perspective from the Complex Network. *Information Sciences*, **627**, 71-96. https://doi.org/10.1016/j.ins.2023.01.084

[4] Cheng, J., Hu, Z., Lu, W., Wang, K. and Cai, Z. (2024) A Hybrid Feature Selection Model for Identifying Groups of Critical Elements in Aero-Engine Assembly. *Quality and Reliability Engineering International*, **40**, 2178-2195. https://doi.org/10.1002/qre.3515

[5] Patel, A. and Singh, B. (2025) Graph Convolutional Network for Structural Equivalent Key Nodes Identification in Complex Networks. *Chaos, Solitons & Fractals*, **196**, Article ID: 116376. https://doi.org/10.1016/j.chaos.2025.116376

[6] Behzadian, M., Otaghsara, S.K., Yazdani, M. and Ignatius, J. (2012) A State-of-the-Art Survey of TOPSIS Applications. *Expert Systems with Applications*, **39**, 13051-13069. https://doi.org/10.1016/j.eswa.2012.05.056

[7] Sun, L. and Lu, P. (2025) TECNN: Identification of Key Nodes in Complex Networks Based on Transformer Encoder and Convolutional Neural Network. *Journal of Computational Science*, **91**, Article ID: 102632. https://doi.org/10.1016/j.jocs.2025.102632

[8] Xiong, Y., Hu, Z., Su, C., Cai, S. and Zhou, T. (2024) Vital Node Identification in Complex Networks Based on Autoencoder and Graph Neural Network. *Applied Soft Computing*, **163**, Article ID: 111895. https://doi.org/10.1016/j.asoc.2024.111895

[9] Veličković, P. (2023) Everything Is Connected: Graph Neural Networks. *Current Opinion in Structural Biology*, **79**, Article ID: 102538. https://doi.org/10.1016/j.sbi.2023.102538

[10] He, Y., Jin, Y., Cao, J., Sui, S., Wang, J. and Ran, B. (2025) Identification of Key Nodes in Urban Bus-Metro Network: A NK-Shell Algorithm Based Neighborhood Ks. *Reliability Engineering & System Safety*, **264**, Article ID: 111448. https://doi.org/10.1016/j.ress.2025.111448

[11] Ye, Q. (2012) Vulnerability Analysis of Rail Transit Based on Complex Network Theory. *China Safety Science Journal*, **22**, 122-126.

[12] Chen, F., Hu, Y.Y. and Li, X.H. (2016) Cascading Failures in Weighted Network of Urban Rail Transit. *Journal of Transportation Systems Engineering and Information Technology*, **16**, 139-145.

[13] Zhang, Y., Shuai, B., Li, S., Zhang, Q. and Zhang, R. (2026) Critical Station Identification Framework Combining Layered Computation with Cross-Layer Fusion for Urban Public Transport Networks. *Reliability Engineering & System Safety*, **274**, Article ID: 112470. https://doi.org/10.1016/j.ress.2026.112470

[14] Wang, L., Zhang, S., Szűcs, G. and Wang, Y. (2024) Identifying the Critical Nodes in Multi-Modal Transportation Network with a Traffic Demand-Based Computational Method. *Reliability Engineering & System Safety*, **244**, Article ID: 109956. https://doi.org/10.1016/j.ress.2024.109956

[15] Lin, Z., Cao, Y., Liu, H., Li, J. and Zhao, S. (2021) Research on Optimization of Urban Public Transport Network Based on Complex Network Theory. *Symmetry*, **13**, Article No. 2436. https://doi.org/10.3390/sym13122436

[16] Zhao, J., Liang, Q. and Ye, Z. (2026) An Intelligent Key Nodes Identification Method in Transportation Networks Based on Gated Attention Multi-Channel GCN. *Eksploatacja i Niezawodność—Maintenance and Reliability*, **28**, Article ID: 218119. https://doi.org/10.17531/ein/218119

[17] Yin, Y., Hu, G., Huang, Z., Chen, J. and Huang, W. (2026) Node Importance Identification Method for Urban Rail Transit Networks Based on Improved Mutual Information Considering Inter-Station Cross-dependencies. *Applied Soft Computing*, **192**, Article ID: 114755. https://doi.org/10.1016/j.asoc.2026.114755

[18] Ministry of Housing and Urban Rural Development of the People's Republic of China, State Administration for Market Regulation (2018) Urban Comprehensive Transport System Planning Standard: GB/T 51328 2018. China Architecture & Building Press.

[19] Wang, L., Wei, H., Guan, Y., Ouyang, L., Xu, D., Han, X., *et al.* (2026) China Public Transport Operation Network Dataset (CPTOND-2025): National-Scale Bus-Metro Vector Dataset. *Scientific Data*, **13**, Article No. 188. https://doi.org/10.1038/s41597-025-06505-4

[20] Seidman, S.B. (1983) Network Structure and Minimum Degree. *Social Networks*, **5**, 269-287. https://doi.org/10.1016/0378-8733(83)90028-x

[21] Hage, P. and Harary, F. (1995) Eccentricity and Centrality in Networks. *Social Networks*, **17**, 57-63. https://doi.org/10.1016/0378-8733(94)00248-9

[22] Burt, R.S. (2004) Structural Holes and Good Ideas. *American Journal of Sociology*, **110**, 349-399. https://doi.org/10.1086/421787

[23] Hirsch, J.E. (2005) An Index to Quantify an Individual's Scientific Research Output. *Proceedings of the National Academy of Sciences*, **102**, 16569-16572. https://doi.org/10.1073/pnas.0507655102

[24] Hamilton, W., Ying, Z.T. and Leskovec, J. (2017) Inductive Representation Learning on Large Graphs. *Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 1024-1034.

[25] Li, Z., Ren, T., Ma, X., Liu, S., Zhang, Y. and Zhou, T. (2019) Identifying Influential Spreaders by Gravity Model. *Scientific Reports*, **9**, Article No. 8387. https://doi.org/10.1038/s41598-019-44930-9

[26] Zuo, Z.Y., Liu, Z.Y. and Yang, G.C. (2025) Key Node Identification of Rail Transit Network Based on Gravity Influence Model. *Journal of Transportation Systems Engineering and Information Technology*, **25**, 102-112.

[27] Joachims, T. (2002). Optimizing Search Engines Using Clickthrough Data. *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Edmonton, 23-26 July 2002, 133-142. https://doi.org/10.1145/775047.775067

[28] Zhang, M., Wang, X., Jin, L., Song, M. and Li, Z. (2022) A New Approach for Evaluating Node Importance in Complex Networks via Deep Learning Methods. *Neurocomputing*, **497**, 13-27. https://doi.org/10.1016/j.neucom.2022.05.010

[29] Albert, R., Jeong, H. and Barabási, A. (2000) Error and Attack Tolerance of Complex Networks. *Nature*, **406**, 378-382. https://doi.org/10.1038/35019019

[30] Kipf, T.N. and Welling, M. (2017) Semi Supervised Classification with Graph Convolutional Networks. *International Conference on Learning Representations*, Tou-

lon, 24-26 April 2017, 1-14.

[31] Guo, H.M., Wang, S.L., Yan, X.F. and Zhang, K. (2024) Node Importance Evaluation Method of Complex Network Based on the Fusion Gravity Model. *Chaos*, *Solitons & Fractals*, **183**, Article ID: 114924. https://doi.org/10.1016/j.chaos.2024.114924

[32] Guan, Y.B., Wang, A.L., Li, B.L. and Xu, R. (2026) A Multiscale Topological Characterization Approach for Identifying Vital Nodes in Complex Networks. *Neurocomputing*, **684**, Article ID: 133516. https://doi.org/10.1016/j.neucom.2026.133516

[33] Guo, H., Yan, X., Liu, Y. and Zhang, J. (2026) A Node Embedded Representation Based Gravity Model for Evaluating the Importance of Nodes in Complex Networks. *Neurocomputing*, **694**, Article ID: 133917. https://doi.org/10.1016/j.neucom.2026.133917