

Article

ReSAN: Relation-Sensitive Graph Representation Learning for Peer Assessment in Educational Scenarios

Xiaoyan Ma ¹, Yujie Fang ², Yongchun Gu ³, Siwei Zhou ³ and Shasha Yang ^{3,*} ¹ School of Educational Sciences, Yili Normal University, Yining 835000, China; 01004@ylnu.edu.cn² School of Computer Science and Technology, Zhejiang Normal University, Jinhua 321004, China; yjfang@zjnu.edu.cn³ Zhejiang Key Laboratory of Intelligent Education Technology and Application, Zhejiang Normal University, Jinhua 321004, China; guyongchun@zjnu.edu.cn (Y.G.); siweizhou@zjnu.edu.cn (S.Z.)

* Correspondence: yangss@zjnu.edu.cn

Abstract

Peer assessment has emerged as a crucial approach for scaling evaluation in educational scenarios, fostering learner engagement, critical thinking, and collaborative learning. Nevertheless, traditional aggregation-based and probabilistic methods often fail to capture the intricate relational dependencies among students and submissions, thereby limiting their capacity to ensure reliable and equitable outcomes. Recent advances in graph neural networks (GNNs) offer promising avenues for representing peer-assessment data as graphs. However, most existing approaches treat all relations uniformly, overlooking variations in the reliability of evaluative interactions. To bridge this gap, we accordingly propose ReSAN (Relation-Sensitive Assessment Network), a novel framework that integrates relation-sensitive attention into the message-passing process. ReSAN dynamically evaluates and weights relationships, enabling the model to distinguish informative signals from noisy or biased assessments. Comprehensive experiments on both synthetic and real-world datasets demonstrate that ReSAN consistently surpasses strong baselines in prediction accuracy and robustness. These findings underscore the importance of explicitly modeling evaluator reliability for effectively capturing the dynamics of peer-assessment networks. Overall, this work advances reliable graph-based evaluation methods and provides new insights into leveraging representation learning techniques for educational analytics.



Academic Editor: Hongbin Pei

Received: 20 October 2025

Revised: 11 November 2025

Accepted: 13 November 2025

Published: 15 November 2025

Citation: Ma, X.; Fang, Y.; Gu, Y.; Zhou, S.; Yang, S. ReSAN: Relation-Sensitive Graph Representation Learning for Peer Assessment in Educational Scenarios. *Mathematics* **2025**, *13*, 3664. <https://doi.org/10.3390/math13223664>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: peer assessment; graph neural networks; educational evaluation; graph representation learning

MSC: 68T07; 68T05; 68T09

1. Introduction

Peer assessment, also known as peer review or peer grading, refers to an educational arrangement in which learners evaluate and specify the quality, value, or performance of work produced by their fellow peers [1]. This mechanism has gained increasing importance in large-scale learning environments, such as MOOCs and blended courses [2,3], where it enables scalable feedback provision while fostering critical thinking, collaboration, and self-reflection [4]. By involving students directly in the evaluation process, peer assessment reduces instructor workload and encourages active engagement, metacognitive awareness, and deeper understanding of subject matter [5]. Its widespread adoption across online and offline educational contexts highlights both its pedagogical value and practical necessity.

Despite these advantages, ensuring the reliability and fairness of peer-generated scores remains a persistent challenge. Peer evaluations are often influenced by graders' varying levels of expertise, subjectivity, or strategic behaviors [6,7]. The tension between scalability and reliability has therefore motivated the development of computational models capable of providing more accurate, robust, and interpretable peer assessment outcomes.

To address the reliability issue, a number of computational methods have been developed to aggregate or refine peer-generated scores [8]. Simple strategies such as averaging or taking the median of peer ratings have been widely used, yet they fail to account for grader reliability or bias [9]. More advanced approaches include probabilistic models such as PG1 [10], PeerRank [11], and Vancouver [12], which model grader ability and attempt to infer latent true scores. Other methods, such as RankwithTA [13], leverage teacher-provided anchor grades to calibrate peer evaluations. While these approaches provide improvements over simple aggregation, they often make restrictive assumptions about graders' consistency or require additional supervision that may not always be available. Moreover, traditional aggregation-based methods typically treat peer assessments as isolated numeric values, overlooking the complex relational structure that exists among students, graders, and submissions [14]. As a result, they may fail to capture higher-order dependencies or different assessment behaviors, limiting their ability to generalize across diverse educational contexts.

Recent advancements in graph neural networks (GNNs) have opened new opportunities for modeling peer assessment data by naturally representing students and submissions as nodes in a relational graph [15]. By leveraging graph-based learning, these approaches can capture dependencies across the peer assessment network and provide richer representations than traditional methods. However, existing graph-based models largely treat all edges equally: they often assume uniform importance across connections or consider only a single type of interaction, thereby overlooking the diversity and specific semantics of assessment relations. In practice, peer assessment networks consist of multiple types of relation, for example, direct grading relations, co-grading relations, or shared course enrollment relations, carrying different levels of information [16].

To bridge this gap, we propose the Relation-Sensitive Assessment Network (ReSAN), a novel framework that dynamically evaluates the relative importance of different relationships in peer assessment graphs. By learning relation-sensitive representations, ReSAN enables more accurate score prediction and a deeper understanding of how diverse assessment signals jointly contribute to reliability.

The main contributions of this paper are threefold:

1. We introduce a relation-sensitive framework that explicitly captures the diversity and distinct characteristics of interactions in peer assessment networks.
2. We design ReSAN, a graph-based model capable of dynamically weighting relationships to enhance representation learning and score prediction.
3. We conduct extensive experiments on both synthetic and real-world datasets, demonstrating that ReSAN outperforms state-of-the-art baselines and improves the reliability of peer grading.

The remainder of this paper is organized as follows. Section 2 reviews related work on peer assessment and graph-based modeling approaches. Section 3 introduces the necessary preliminaries and formulates the graph neural network model. Section 4 presents the proposed ReSAN framework in detail. Section 5 reports experimental settings and results on benchmark datasets. Section 6 discusses findings, implications, and potential extensions. Finally, Section 7 concludes the paper and outlines future research directions.

2. Related Work

The rapid expansion of online learning platforms, massive open online courses (MOOCs), and technology-enhanced classrooms has fundamentally reshaped the delivery of education [17–19]. With increasing class sizes and growing demand for personalized feedback, traditional instructor-led grading has become infeasible at scale. Peer assessment has emerged as a practical and pedagogically valuable solution to this scalability challenge [20]. By involving students directly in the evaluation process, peer assessment not only distributes grading workload but also promotes deeper cognitive engagement, critical reflection, and collaborative learning. Recent surveys in educational technology have highlighted its ability to cultivate transferable skills such as self-regulation and constructive feedback [21], making it a cornerstone in large-scale digital education. Nonetheless, achieving both scalability and reliability remains difficult: while peer assessment enables broader participation, the variability in graders' expertise, motivation, and biases can compromise fairness and accuracy [8,22,23]. This duality has motivated a wide range of computational methods to enhance the robustness and trustworthiness of peer-generated evaluations.

2.1. Peer Assessment in Educational Scenarios

Early studies on peer assessment primarily relied on straightforward aggregation techniques, such as computing the mean or median of peer ratings. While computationally efficient, these approaches fail to account for differences in grader reliability and are vulnerable to noisy evaluations, often resulting in inaccurate estimations of submission quality. To address these limitations, more sophisticated approaches were proposed. Probabilistic models such as PG1 [10] use Bayesian inference to jointly estimate grader reliability and the latent true scores of assignments, iteratively refining predictions. PeerRank [11] employs a PageRank-like propagation mechanism to weight student evaluations by grader reliability, while Vancouver [12] estimates grader consistency and applies iterative reweighting to correct scores. Teacher-assisted methods like RankwithTA [13] incorporate anchor grades provided by instructors to calibrate peer assessments. These strategies significantly improve reliability compared to simple aggregation, yet they often rest on rigid assumptions (e.g., stable grader ability, Gaussian noise) and can require additional supervisory input that may not always be feasible. More importantly, they tend to treat peer scores as isolated numeric signals rather than relational data, failing to exploit the structural dependencies inherent in peer assessment networks. More recent work has explored filtering or weighting raters based on behavioral and linguistic signals, such as pruning MOOC raters by score variance and lexical diversity to improve the reliability of peer assessment [24].

2.2. Graph-Based Modeling of Peer Assessment

The relational nature of peer assessment naturally lends itself to graph-based representation, where students and submissions form nodes and grading interactions define edges. This perspective has opened the door for graph neural networks (GNNs) to be applied in educational settings. Foundational models such as the Graph Convolutional Network (GCN) [25] and Graph Attention Network (GAT) [26], as well as inductive variants like GraphSAGE [27], have demonstrated the ability to aggregate information from neighbors and capture higher-order dependencies. Sheaf Neural Networks [28] assign local vector spaces to nodes and edges, enabling more expressive representations of relational structures. Framelet-based GNNs [29] leverage multi-scale spectral decompositions to capture both low-frequency and high-frequency information in graphs. Their success in domains ranging from social networks [30,31] to biological systems [32,33] suggests strong potential for educational data. Indeed, early applications of GNNs to peer assessment, such as GCN-SOAN [34], showed that graph-based representations outperform traditional baselines by

leveraging the structural properties of grading networks. These developments collectively highlight the promise of graph-based approaches for capturing complex dependencies in peer assessment networks.

2.3. Relation-Aware GNNs in Broader Graph Learning

The broader GNN literature has increasingly emphasized the need for relation sensitivity and edge-awareness. FraS-HNN [35] innovatively achieves educational relationship modeling by constructing a signed hypergraph network between students and questions. EduCross [36] utilizes dual-channel adversarial bipartite hypergraph learning to focus on constructing high-order relational networks among educational content modalities. Relational GCN (R-GCN) [37] and CompGCN [38] extend standard GCNs by introducing relation-specific parameters or embeddings, enabling them to capture the semantics of multi-relational graphs. Graph Isomorphism Network (GIN) [39] and more recent GATv2 [40] further push expressive power, with attention mechanisms that dynamically assign importance to neighbors, thereby improving flexibility and interpretability. In parallel, heterogeneous graph neural networks [41] leverage type-specific aggregators and semantic attention to integrate information across multiple node and edge types. These advances have yielded state-of-the-art performance in areas such as knowledge graph reasoning [42], recommendation systems [43,44], and molecular property prediction [45], demonstrating the effectiveness of explicitly modeling relational diversity. Despite these achievements, their adaptation to educational domains remains limited. Peer assessment networks are characterized by nuanced relation types [46], sparse interactions [47], and evolving grading behaviors—conditions [48] under which generic relation-aware GNNs may not directly transfer. This gap resonates with a broader trend in graph learning, where increasing attention has been devoted to modeling relation sensitivity and dynamic edge importance. Recent advances in this direction highlight the importance of adaptively assigning weights to different relational signals. Attention-based architectures exemplify this trend by dynamically evaluating the contribution of neighbors, thereby improving robustness in noisy or imbalanced networks. In the context of peer assessment, such mechanisms are particularly relevant, as graders differ in reliability, expertise, and engagement. Nonetheless, existing approaches have not systematically exploited this property. To address this gap, we introduce ReSAN, which integrates structured relational modeling with relation-sensitive attention to dynamically capture the varying importance of peer assessment interactions.

In summary, prior research on peer assessment has demonstrated the pedagogical benefits of peer evaluation and developed a range of statistical and probabilistic models to address grading bias and fairness. Meanwhile, GNNs have opened new opportunities for leveraging relational structures in peer-assessment networks. However, existing methods either ignore structural dependencies (statistical models) or fail to differentiate between different relations (standard GNNs). To bridge this gap, we propose ReSAN, a relation-sensitive graph representation learning framework that dynamically evaluates and weights assessment relations, enabling more reliable and robust peer-assessment outcomes.

Remark 1. *The term uniform-edge limitation in this study specifically refers to the assumption adopted by most peer-assessment-specific GNN models that treat all evaluation edges as equally informative. This constraint overlooks the heterogeneity of evaluator reliability and thus limits the model's ability to capture relational nuances. Our proposed ReSAN addresses this issue by introducing relation-sensitive attention to differentiate interaction strengths dynamically.*

3. Preliminaries and Notation

3.1. Graph Neural Networks

A graph is represented as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the vertex set with size $N = |\mathcal{V}|$, and $\mathcal{E}$ is the edge set with size $M = |\mathcal{E}|$. Each vertex $v \in \mathcal{V}$ is associated with a feature vector, and we denote the vertex feature matrix as $X \in \mathbb{R}^{N \times d}$, where d is the feature dimension. The graph structure is encoded by an adjacency matrix $A \in \{0, 1\}^{N \times N}$, where

$$A(u, v) = \begin{cases} 1, & \text{if } (u, v) \in \mathcal{E}, \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

The degree of a node v is defined as $D(v) = \sum_{u \in \mathcal{V}} A(u, v)$, and the degree matrix is denoted as $D \in \mathbb{R}^{N \times N}$.

In [25,49], the classical form of a graph convolutional layer $f(X, \Theta)$ is defined as:

$$X^{(\ell+1)} = \sigma(\hat{D}^{-1/2} \hat{A} \hat{D}^{-1/2} X^{(\ell)} \Theta^{(\ell)}), \quad (2)$$

where $\hat{A} = A + I$ is the adjacency matrix with self-loops, $\hat{D}$ is the corresponding degree matrix, $X^{(\ell)} \in \mathbb{R}^{N \times C}$ denotes the vertex representation at layer ℓ with $X^{(0)} = X$, $\Theta^{(\ell)}$ is the trainable weight matrix, and $\sigma(\cdot)$ denotes a nonlinear activation function.

This formulation follows the message-passing paradigm, where each node updates its representation by aggregating information from its neighbors, enabling GNNs to capture both local and higher-order dependencies in the graph.

3.2. Social–Ownership–Assessment Network

We consider a setting in which a group of n users, denoted by $\mathcal{U}$ (e.g., students or reviewers), are responsible for evaluating a set of m items $\mathcal{I}$ (e.g., assignments or scholarly articles). Each item $i \in \mathcal{I}$ is assumed to possess a latent ground-truth value $v_i \in \mathbb{R}^+$, which may correspond, for instance, to an expert's judgment.

The evaluation behavior of users is represented by an assessment matrix

$$A = [a_{ui}] \in \mathbb{R}^{n \times m},$$

where a_{ui} denotes the score assigned by user $u \in \mathcal{U}$ to item $i \in \mathcal{I}$. If user u does not evaluate item i , we set $a_{ui} = 0$; otherwise, $a_{ui} > 0$. In practice, only a small subset of user–item pairs contain evaluations, rendering the assessment matrix A highly sparse. Equivalently, the same data can be represented as a weighted bipartite graph consisting of user nodes, item nodes, and weighted edges that encode assessment relationships.

1. **Social relations:** Social relations capture the interpersonal connections among users in $\mathcal{U}$, such as friendships, collaborations, or potential conflicts of interest. These relations can influence evaluation behavior by introducing biases toward friends or competitors and provide important context for interpreting assessment scores.
2. **Ownership relations:** Ownership relations encode the authorship or contribution of users to items in $\mathcal{I}$. Multiple users may contribute to a single item, forming a many-to-one relationship. This type of relation allows the model to distinguish between self-assessment (a user evaluating their own work) and peer-assessment, while also supporting flexible modeling of group contributions.
3. **Assessment relations:** Assessment relations represent the actual evaluations performed by users, as captured in the assessment matrix A . Each directed edge (u, i) with weight a_{ui} reflects the score assigned by user u to item i , including both self-assessment and peer-assessment. Explicitly representing these relations enables the framework to analyze the reliability and patterns of evaluation across the network.

The Social–Ownership–Assessment (SOAN) framework provides a unified representation of peer assessment by jointly modeling three key relational types: social connections among users, ownership of items, and the assessment interactions themselves, as shown in Figure 1. For an assignment that may be collaboratively completed by several students, both self-assessment and peer-assessment are involved. Beyond these explicit evaluation links, the reliability of peer-assessment is also shaped by hidden relational factors such as conflicts of interest or evaluator bias. By explicitly incorporating social and ownership relations, SOAN captures contextual information that can influence evaluation behavior, such as biases, conflicts of interest, and group contributions. This enriched relational modeling allows downstream learning methods, such as graph neural networks, to leverage structural dependencies in the network, leading to more accurate and interpretable prediction of ground-truth evaluations compared with approaches that consider only the user–item assessment matrix.

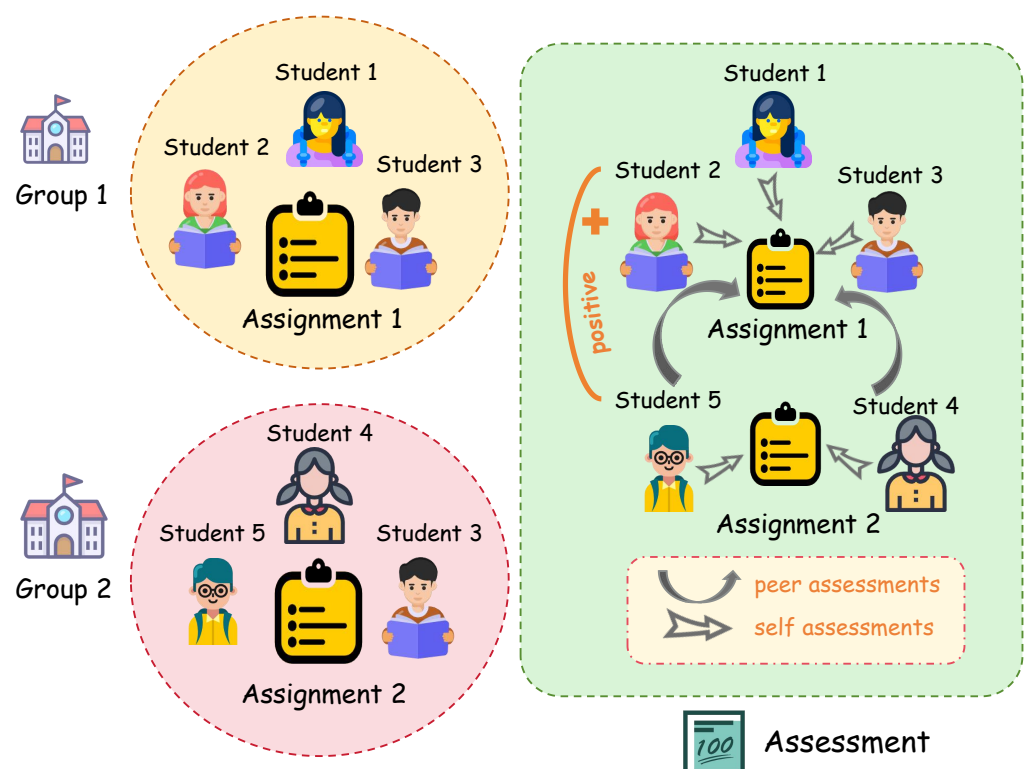


Figure 1. Illustration of diverse relational types in peer assessment scenarios.

4. Proposed Model: ReSAN

4.1. Framework Overview

ReSAN is a relation-sensitive assessment network designed to enhance the reliability of peer assessment by jointly modeling multiple relational perspectives. As illustrated in Figure 2, the model operates on the Social–Ownership–Assessment Network (SOAN) by first integrating social connections, ownership links, and assessment scores into a unified relational graph that simultaneously represents users and items. Initial embeddings are assigned to nodes and iteratively refined through a relation-sensitive attention propagation mechanism. This mechanism adaptively weighs different relational signals, highlighting reliable evaluators while suppressing noise or bias. After several propagation layers, enriched item embeddings are obtained and mapped through a regression head to produce calibrated score predictions. Training is conducted in a semi-supervised manner, where ground-truth labels are available for only a subset of items.

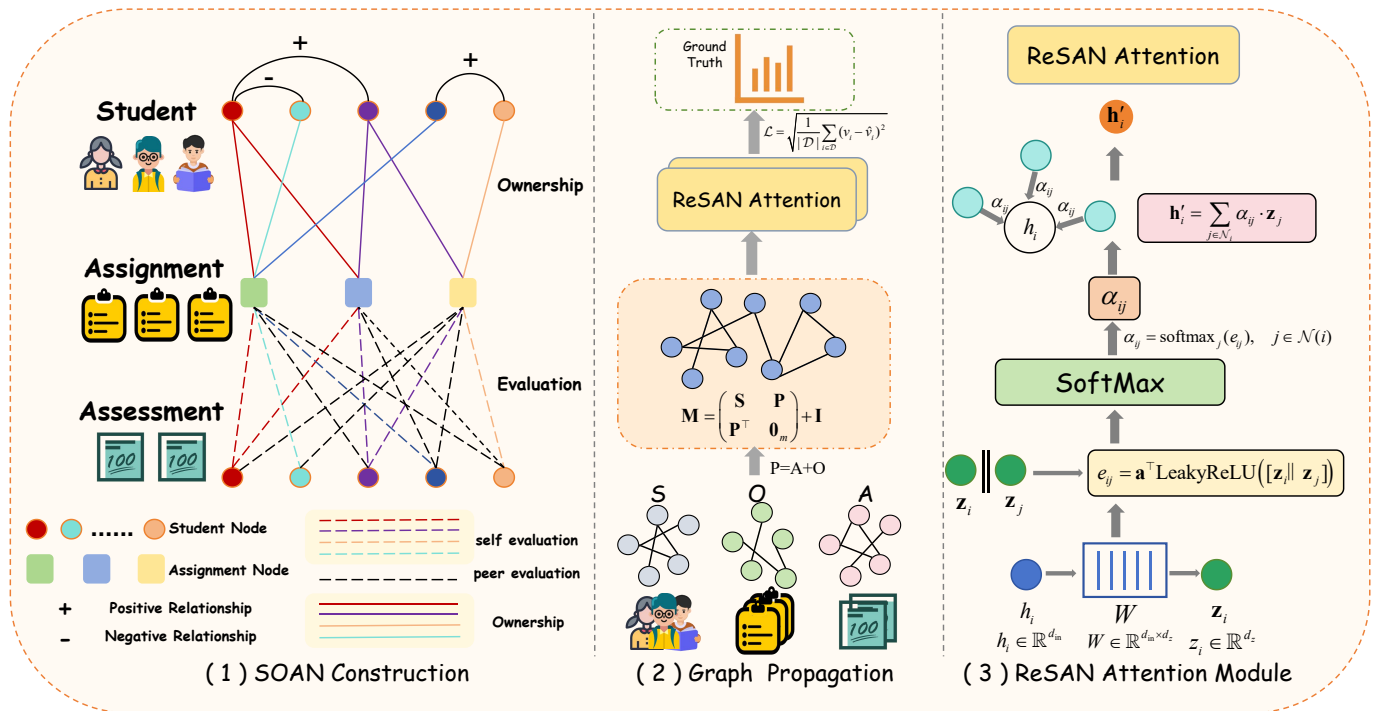


Figure 2. An overview of ReSAN.

4.2. Relation-Sensitive Representation Learning

Building upon the Social–Ownership–Assessment Network (SOAN), we develop a relation-sensitive representation learning framework for peer assessment. The proposed methodology comprises three key stages: (i) constructing a unified relational matrix that captures multiple dimensions of relationships, (ii) performing relation-sensitive aggregation to learn informative embeddings, and (iii) employing semi-supervised learning to predict ground-truth scores.

Given users $\mathcal{U}$ and items $\mathcal{I}$, we recall three relation types: social links among users, ownership links between users and items, and assessment links capturing evaluations. These relations are encoded by adjacency matrices $S \in \mathbb{R}^{n \times n}$, $O \in \mathbb{R}^{n \times m}$, and $A \in \mathbb{R}^{n \times m}$, respectively. To unify them, we form a block matrix

$$M = \begin{bmatrix} S & N \\ N^T & 0_{m \times m} \end{bmatrix} + I, N = O + A \quad (3)$$

where I is the identity matrix ensuring self-loops, and $0_{m \times m}$ denotes a zero block for item–item connections. The resulting $M \in \mathbb{R}^{(n+m) \times (n+m)}$ encodes the SOAN structure with multiple types of relationships in a form suitable for neural message passing.

Let $h_i^{(\ell)} \in \mathbb{R}^d$ denote the embedding of node i at layer ℓ . For each neighbor $j \in \mathcal{N}(i)$ defined by M , we compute a relation-sensitive attention score

$$e_{ij}^{(\ell)} = a^T \text{LeakyReLU}(W^{(\ell)} [h_i^{(\ell)} \| h_j^{(\ell)}]), \quad (4)$$

where $W^{(\ell)} \in \mathbb{R}^{d' \times 2d}$ is a learnable projection, $a \in \mathbb{R}^{d'}$ is an attention vector, and $[\cdot \| \cdot]$ denotes concatenation. The normalized attention coefficient is obtained by

$$\alpha_{ij}^{(\ell)} = \frac{\exp(e_{ij}^{(\ell)})}{\sum_{k \in \mathcal{N}(i)} \exp(e_{ik}^{(\ell)})}. \quad (5)$$

The embedding is updated as

$$h_i^{(\ell+1)} = \sigma \left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij}^{(\ell)} W^{(\ell)} h_j^{(\ell)} \right), \quad (6)$$

where $\sigma(\cdot)$ is a nonlinear activation such as ELU. Stacking L layers allows the model to capture higher-order dependencies across social, ownership, and assessment relations.

For each item $i \in \mathcal{I}$, we obtain the final embedding $h_i^{(L)}$ and predict its ground-truth score by a regression head:

$$\hat{v}_i = w^\top h_i^{(L)} + b, \quad (7)$$

where w and b are learnable parameters. The model is trained by minimizing the root mean squared error (RMSE):

$$\mathcal{L} = \sqrt{\frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} (v_i - \hat{v}_i)^2}, \quad (8)$$

where $\mathcal{D}$ denotes the set of items with known ground-truth scores.

The architecture of the proposed attention mechanism is illustrated in Figure 3. Student 1 completes an assignment and first provides a self-assessment. Subsequently, Students 3, 4, and 5 also evaluate the same submission. However, their assessments are influenced by underlying social relations. Specifically, Student 4 has a negative relationship with Student 1, which leads to a biased evaluation with a lower attention weight (e.g., $\alpha = 0.2$). In contrast, Student 5 maintains a close positive relation with Student 1, resulting in a relatively higher weight (e.g., $\alpha = 0.3$). Meanwhile, Student 3 has no significant social ties with Student 1, and thus their evaluation is considered more neutral and assigned the highest reference weight (e.g., $\alpha = 1$). This differentiated attention allocation enables the model to dynamically adjust the importance of peer ratings, thereby enhancing the reliability and fairness of the aggregated assessment.

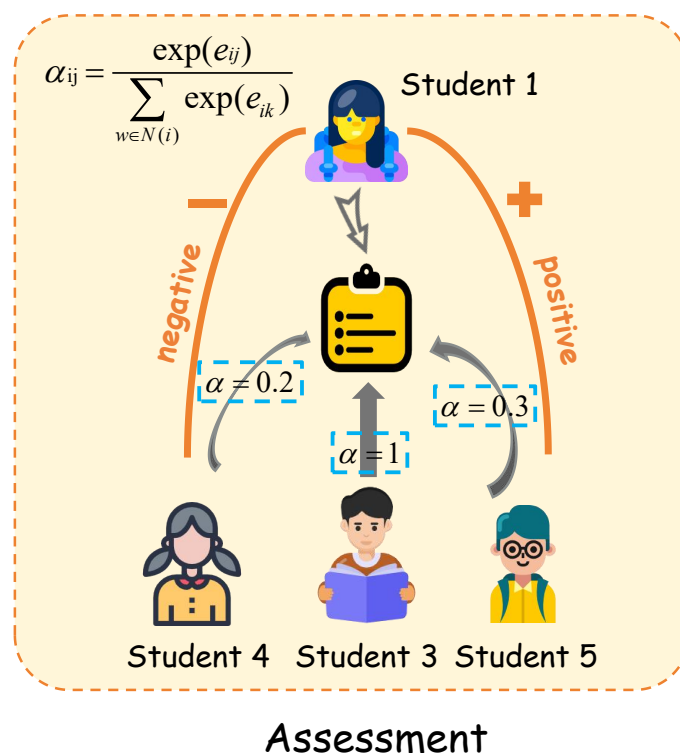


Figure 3. Illustration of the attention mechanism in peer assessment.

4.3. Algorithm Implementation

The pseudocode description of ReSAN is presented in the following Algorithm 1.

Algorithm 1 ReSAN: Relation-Sensitive Assessment Network

Input: User set $\mathcal{U}$, item set $\mathcal{I}$, relation set $\mathcal{R}$, assessment matrix $A \in \mathbb{R}^{n \times m}$, initial node features $\mathbf{X}^{(0)}$, number of layers L

Output: Predicted scores $\hat{v}_i$ for each item $i \in \mathcal{I}$

1. Graph construction: Integrate $\mathcal{E}_{\text{soc}}$, $\mathcal{E}_{\text{own}}$, and $\mathcal{E}_{\text{ass}}$ into graph $\mathcal{G}$.

2. Initialization: Assign node embeddings $\mathbf{h}_i^{(0)}$ for all $i \in \mathcal{V}$ from $\mathbf{X}^{(0)}$.

For $\ell = 0, 1, \dots, L - 1$ do:

 Message encoding: For each edge $(i, j) \in \mathcal{E}$, compute the raw attention score

$$e_{ij}^{(\ell)} = \mathbf{a}^\top \text{LeakyReLU}\left(W^{(\ell)} [\mathbf{h}_i^{(\ell)} \parallel \mathbf{h}_j^{(\ell)}]\right).$$

 Attention normalization: Apply softmax over neighbors:

$$\alpha_{ij}^{(\ell)} = \frac{\exp(e_{ij}^{(\ell)})}{\sum_{k \in \mathcal{N}(i)} \exp(e_{ik}^{(\ell)})}.$$

 Node update: Update each node embedding:

$$\mathbf{h}_i^{(\ell+1)} = \sigma \left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij}^{(\ell)} W^{(\ell)} \mathbf{h}_j^{(\ell)} \right).$$

Prediction: Obtain item-level outputs $\hat{v}_i = w^\top \mathbf{h}_i^{(L)} + b$, where w, b are learnable parameters.

Optimization: Train the model by minimizing the root mean squared error (RMSE):

$$\mathcal{L} = \sqrt{\frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} (v_i - \hat{v}_i)^2}.$$

Remark 2. The construction of M integrates multiple types of relations into a single structure, while the attention-based aggregation dynamically evaluates the relative importance of each relation for every prediction. Intuitively, the model learns when to trust social ties, when to emphasize authorship, and when to rely on assessment signals, thereby offering a unified yet flexible framework for peer assessment.

Remark 3. In contrast to conventional attention mechanisms that impose restrictive linear forms on neighbor importance, our relation-sensitive design employs a more expressive scoring function. Specifically, the attention score is derived from a non-linear transformation of both the target node and its neighbors, allowing interactions to be captured beyond simple additive or dot-product forms. This expressiveness enables the model to recognize subtle differences in relation patterns, such as distinguishing cases where two neighbors provide complementary signals from those where they provide redundant information, thereby dynamically adapting the relative contribution of each relation to the prediction context.

5. Experiments

This section provides a systematic evaluation of the ReSAN model for ground-truth prediction in peer assessment. The analysis examines its accuracy, robustness, and practical utility across diverse conditions, including real-world, complex, and adversarial scenarios, using both real and synthetic datasets. Section 5.1 introduces the experimental

setup, including the datasets employed and baseline methods for comparison. Section 5.2 presents results on real-world datasets, demonstrating the model’s performance in authentic educational settings. Section 5.3 reports results on synthetic datasets, where controlled experiments provide additional insights into the model’s behavior under varying conditions. Section 5.4 investigates the model’s behavior under strategic assessment settings, highlighting its resilience to manipulative grading strategies. Section 5.5 provides parameter sensitivity analysis to understand the influence of key hyperparameters, and Section 5.6 conducts ablation studies to disentangle the contributions of different model components.

5.1. Experimental Setup

5.1.1. Datasets

We employ the peer grading dataset collected by [9], which records both TA grades and student-provided self/peer grades across multiple exercise sheets. Comprehensive details of the real-world dataset are presented in Table 1. The dataset comprises 219 students distributed across 79 study groups. For each assignment, students provided self-assessments based on the reference solution and conducted blind evaluations of two submissions from other groups. Furthermore, each submission was independently graded by a teaching assistant following conventional procedures, thereby producing three types of scores: self-assessments, peer assessments, and ground-truth evaluations.

Table 1. Summary statistics of real-world peer grading datasets.

Statistic	Asst. 1	Asst. 2	Asst. 3	Asst. 4
Average Grades				
Ground-truth	0.62 ± 0.27	0.71 ± 0.24	0.69 ± 0.33	0.59 ± 0.27
Peer	0.70 ± 0.26	0.76 ± 0.23	0.75 ± 0.31	0.68 ± 0.29
Self	0.74 ± 0.22	0.80 ± 0.22	0.82 ± 0.26	0.76 ± 0.24
Number of				
Exercises	3	4	5	3
Groups	75	77	76	79
Students	183	206	193	191
Items	225	308	380	237
Peer grades	965	1620	1889	1133
Self grades	469	755	890	531

We introduce the models employed for generating the synthetic datasets, in accordance with the approach described in [34].

- Ground-truth valuation. Each item $i \in A$ is assigned a latent quality value v_i , drawn from a mixture of two Gaussian distributions:

$$v_i \sim \sum_{c=1}^2 \pi_c \mathcal{N}(x; \mu_c, \sigma_c), \quad (9)$$

where the parameters of the mixture are specified by the weights $\pi = (\pi_1, \pi_2)$, means $\mu = (\mu_1, \mu_2)$, and standard deviations $\sigma = (\sigma_1, \sigma_2)$. The tuple (π_k, μ_k, σ_k) fully characterizes the k -th component.

- Social network. We simulate interpersonal connections among users using the Erdős–Rényi random graph $G(n, p)$. In this model, each of the $\binom{n}{2}$ possible user pairs forms an edge independently with probability p .
- Ownership network. To capture the individual responsibility of submissions, every user is randomly matched to exactly one item, creating a bijective relationship between

users and their owned submissions. This setup is consistent with prior works on peer assessment that do not allow group ownership.

- **Assessment network.** For every item $i \in A$, a set of k graders is chosen uniformly at random, denoted as $N(i) \subset U$ with $|N(i)| = k$. Each grader $u \in N(i)$ then assigns a grade A_{ui} to item i , generated under one of the following mechanisms:
- **Strategic model.**

$$A_{ui} = \begin{cases} 1, & \text{if grader } u \text{ is a friend of the owner } j \text{ of item } i \text{ } (s_{uj} \cdot o_{ji} = 1), \\ \mathcal{N}(v_i, \sigma_H), & \text{otherwise.} \end{cases} \quad (10)$$

This reflects the behavior where friends collude by awarding the maximum score to each other, while remaining reasonably fair in evaluating unrelated peers.

- **Bias–reliability model.**

$$A_{ui} \sim \mathcal{N}(\hat{\mu}, \hat{\sigma}), \quad \hat{\mu} = v_i + \alpha, \quad \hat{\sigma} = \sigma_{\max}(1 - \beta v_l), \quad (11)$$

where v_l denotes the true valuation of the item owned by user u . The bias parameter $\alpha \in [-1, 1]$ adjusts for generosity ($\alpha > 0$) or strictness ($\alpha < 0$), while the reliability parameter β determines the dependence of grading reliability on the grader's own item quality. Here, $\sigma_{\max}$ represents the maximum possible standard deviation, corresponding to the least reliable case.

To facilitate a clearer understanding of the data generation process, we summarize the key components and their mathematical formulations in Table 2.

Table 2. Synthetic data generation process.

Component	Formulation
Ground-truth valuation	$v_i \sim \sum_{c=1}^2 \pi_c \mathcal{N}(x; \mu_c, \sigma_c), \quad \pi = (\pi_1, \pi_2), \mu = (\mu_1, \mu_2), \sigma = (\sigma_1, \sigma_2)$
Social network	$G(n, p), \quad \Pr[(u, v) \in E] = p$
Ownership network	One-to-one mapping: $u \leftrightarrow i$
Assessment network	For $i \in A$, $N(i) \subset U$, $ N(i) = k$
Strategic model	$A_{ui} = \begin{cases} 1, & s_{uj} \cdot o_{ji} = 1, \\ \mathcal{N}(v_i, \sigma_H), & \text{otherwise,} \end{cases}$
Bias–reliability model	$A_{ui} \sim \mathcal{N}(\hat{\mu}, \hat{\sigma}), \quad \hat{\mu} = v_i + \alpha, \quad \hat{\sigma} = \sigma_{\max}(1 - \beta v_l)$

5.1.2. Baselines

Our ReSAN model is evaluated against a range of baseline methods:

- **PeerRank [11]:** PeerRank is inspired by the PageRank algorithm, modeling peer assessments as a graph structure. It iteratively computes a “reputation” score within the peer network to adjust and weight students’ final grades.
- **PG1 [10]:** PG1 applies statistical techniques to correct grading bias among students. It standardizes or adjusts individual scores during aggregation to improve fairness and consistency of the final outcomes.
- **RankwithTA [13]:** RankwithTA incorporates Teaching Assistant (TA) grades as anchor points into the peer assessment process. By leveraging TA evaluations, it reduces noise and instability, thus enhancing the reliability of rankings.
- **Vancouver [12]:** The Vancouver method employs Bayesian inference to model students’ grading behavior. It simultaneously estimates the true quality of submissions and the grading ability of students, mitigating subjectivity-related bias.

- GCN-SOAN [14]: GCN-SOAN proposes a GNN-based general peer assessment framework that captures complex behaviors by modeling the evaluation system as a multi-relational network.
- Average & Median: Since models such as PeerRank, PG1, and RankwithTA treat users and items interchangeably, they are not directly applicable to our dataset, which contains individual evaluations of group submissions. To enable comparison, we aggregate individual scores by averaging the grades assigned by all members within each group.

For all baseline models, we adopt either the optimal parameters recommended in their original publications or obtain competitive configurations through parameter search.

5.1.3. Experimental Settings

The proposed model is implemented using PyTorch 2.9.1 Geometric library. In all experiments, we employ a two-layer network architecture. Given the limited scale of the datasets, we adopt the Monte Carlo [50] cross-validation strategy for evaluation. For real-world datasets, 20% of the data is used for training and 80% for testing, whereas for synthetic datasets, 10% is used for training and 90% for testing, following the same data split strategy as GCN-SOAN [14]. To ensure robustness, each experiment is repeated four times, with 800 training epochs per run. The final performance is reported as the average root mean square error (RMSE) across all runs. The hyperparameter search space for the model is summarized in Table 3.

Table 3. Hyperparameter searching space for ground-truth evaluation.

Hyperparameters	Searching Space
Learning rate	$1 \times 10^{-1}, 1 \times 10^{-2}, 1 \times 10^{-3}, 2 \times 10^{-3}, 3 \times 10^{-3}, 4 \times 10^{-3}, 5 \times 10^{-3}$
Weight decay	$0, 1 \times 10^{-5}$
Hidden Size	32, 64, 128, 256, 512
Dropout ratio	0.1, 0.2, 0.3, 0.4, 0.5
Heads	1, 2, 3, 4, 5, 6

5.2. Results on Real-World Datasets

Our proposed model, ReSAN, demonstrates a consistent and superior performance across all experimental settings, as evidenced by the results presented in Table 4. The evaluation metric employed here is the Root Mean Square Error (RMSE), which quantifies the difference between the model-predicted grades and the ground-truth scores. A lower RMSE value indicates a higher accuracy in grading alignment. Under both the peer evaluation and the combined peer and self evaluation conditions, ReSAN achieves the lowest RMSE values on all four assignments, indicating its effectiveness in reducing grading discrepancy compared to all baseline models.

In the peer evaluation setup, ReSAN attains RMSE scores of 0.1690, 0.1617, 0.1835, and 0.1749 on the four assignments, respectively, outperforming the next best baseline in each case. Similarly, under the peer and self evaluation condition, it yields scores of 0.1668, 0.1556, 0.1829, and 0.1747, again surpassing all other methods. This consistent improvement highlights the robustness and generalizability of our approach across different evaluation configurations and assignment types.

Notably, ReSAN also exhibits a more stable performance profile compared to models such as RankwithTA and Vancouver, which show greater variability in their RMSE across assignments. The ability of our method to consistently achieve the lowest error rates underscores the effectiveness of the proposed architecture in capturing meaningful patterns

in peer assessment data, potentially through its enhanced representation learning and attention mechanisms.

Table 4. Root mean square errors of eight methods on two real-world datasets. Red, yellow, and blue indicate the first, second, and third best results, respectively.

Model	Peer Evaluation			
	Asst. 1	Asst. 2	Asst. 3	Asst. 4
Average	0.1917	0.1712	0.1902	0.1989
Median	0.1991	0.1843	0.2047	0.2250
PeerRank	0.1913	0.1762	0.2235	0.2087
PG1	0.1919	0.1669	0.2110	0.2161
RankwithTA	0.1922	0.1903	0.2183	0.1740
Vancouver	0.1851	0.1688	0.1951	0.2071
GCN-SOAN	0.1795	0.1673	0.1869	0.1822
ReSAN (Ours)	0.1690	0.1617	0.1835	0.1749

Model	Peer and Self Evaluation			
	Asst. 1	Asst. 2	Asst. 3	Asst. 4
Average	0.1944	0.1681	0.2023	0.2117
Median	0.2111	0.1750	0.2333	0.2538
PeerRank	0.1888	0.1721	0.2203	0.2168
PG1	0.2009	0.1680	0.2111	0.2304
RankwithTA	0.1884	0.1845	0.2137	0.1792
Vancouver	0.1815	0.1672	0.1945	0.2101
GCN-SOAN	0.1778	0.1621	0.1840	0.1821
ReSAN (Ours)	0.1668	0.1556	0.1829	0.1747

5.3. Results on Synthetic Datasets with Bias-Reliability

We design synthetic experiments under a controlled protocol to evaluate the robustness of ReSAN. A total of $n = 500$ users and $m = 500$ items are considered, with a random one-to-one ownership assignment.

For ground-truth generation, true item scores are sampled from a bimodal Gaussian mixture distribution with parameters $\mu = (0.3, 0.7)$, $\sigma = (0.1, 0.1)$, and $\pi = (0.2, 0.8)$. For assessment network generation, each item receives $k = 3$ peer grades. Unless otherwise specified, the maximum grading variance is fixed at $\sigma_{\max} = 0.25$, and both the bias parameter α and reliability parameter β are set to zero, reflecting a neutral evaluation environment without systematic distortions. For social network generation, no user–user social relations are incorporated in the default setup. To examine the impact of individual factors, we vary one parameter (e.g., k , α , μ , or β) while holding all others at their default values. Model performance is reported in terms of root mean squared error (RMSE) between predicted and ground-truth scores.

Figure 4a shows how the number of evaluators k affects model performance. As the number of evaluators increases, the error of all models decreases, indicating that more peer assessments improve grading accuracy. ReSAN consistently achieves the lowest error across all settings, with a particularly clear advantage when the number of evaluators is small. This shows that ReSAN can extract reliable information even from a small number of peer reviews, reducing the grading workload while maintaining accurate results.

Figure 4b shows the impact of evaluator bias α . As α increases, representing more lenient or stricter evaluators, the errors of baseline methods increase noticeably, while ReSAN remains relatively stable. This indicates that ReSAN can account for overall grading tendencies and handle biased evaluations effectively.

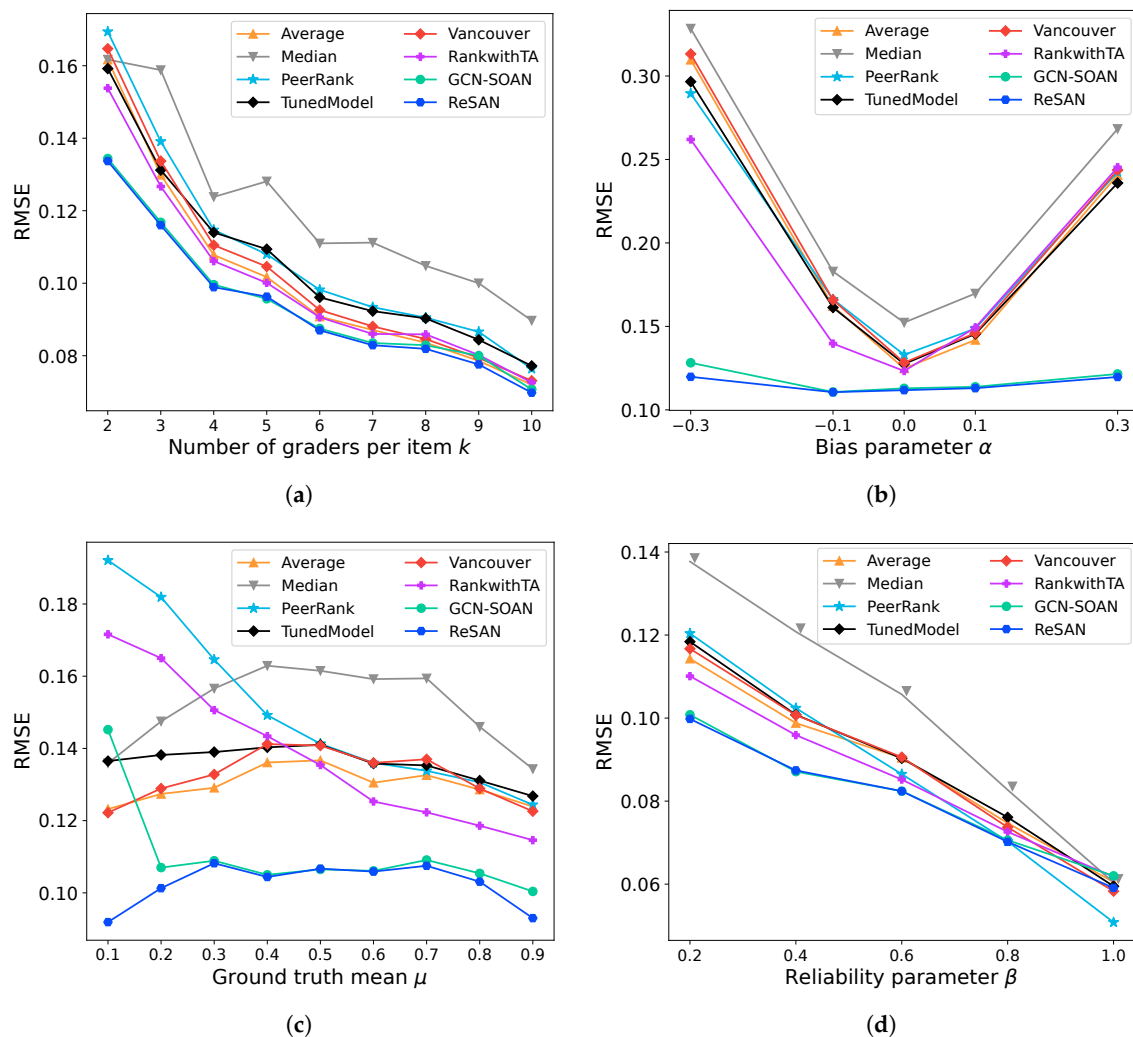


Figure 4. Results of controlled experiments on synthetic datasets, showing the effects of varying key parameters on model performance measured by RMSE. (a) Effect of the number of graders per item k ; (b) Effect of the bias parameter α ; (c) Effect of the ground truth mean μ ; (d) Effect of the reliability parameter β .

Figure 4c reports results for different ground-truth score distributions. When the distribution changes from the default bimodal form to skewed or uniform forms, all models show some performance drop. However, ReSAN maintains a clear advantage, especially for skewed distributions, suggesting that it can make good use of the relationships among evaluators to maintain reliable grading even when the data distribution is uneven.

Figure 4d examines the effect of evaluator reliability correlation β . Baseline methods show inconsistent changes in error, while ReSAN consistently benefits from the alignment between evaluator reliability and grading accuracy. This shows that ReSAN can use reliability information among evaluators to improve grading accuracy and stability across different conditions.

Overall, these experiments show that ReSAN performs well under different numbers of evaluators, bias levels, score distributions, and reliability conditions. Its adaptive relation-sensitive mechanism allows it to extract useful information from peer assessments, making it robust and practical for a variety of educational settings.

5.4. Results on Synthetic Datasets with Strategic Assessment

Building upon the default synthetic setup described in Section 5.3, we introduce strategic behaviors among users to examine the robustness of peer-assessment methods. In this scenario, a social network is generated among the $n = 500$ users using an Erdős–Rényi random graph model with edge probability p . The peer assessment network is subsequently constructed according to the strategic model, while all other parameters, including the number of items m , the ground truth score distribution, the number of peer evaluations k , and $\sigma_{\max}$ remain identical to the default setting.

To assess the impact of collusive behavior, we systematically vary the connection probability p while holding other parameters constant, and evaluate model performance using the RMSE between predicted and ground-truth scores. As shown in Figure 5, ReSAN consistently attains lower RMSE than baseline methods across a wide range of connection probabilities. Its advantage becomes increasingly pronounced as the density of colluding social ties rises, indicating that the model is robust to biased or coordinated grading behaviors. These findings demonstrate that ReSAN effectively leverages relation-sensitive attention mechanisms to differentiate between reliable and unreliable evaluations, thereby mitigating the adverse effects of strategic grading.

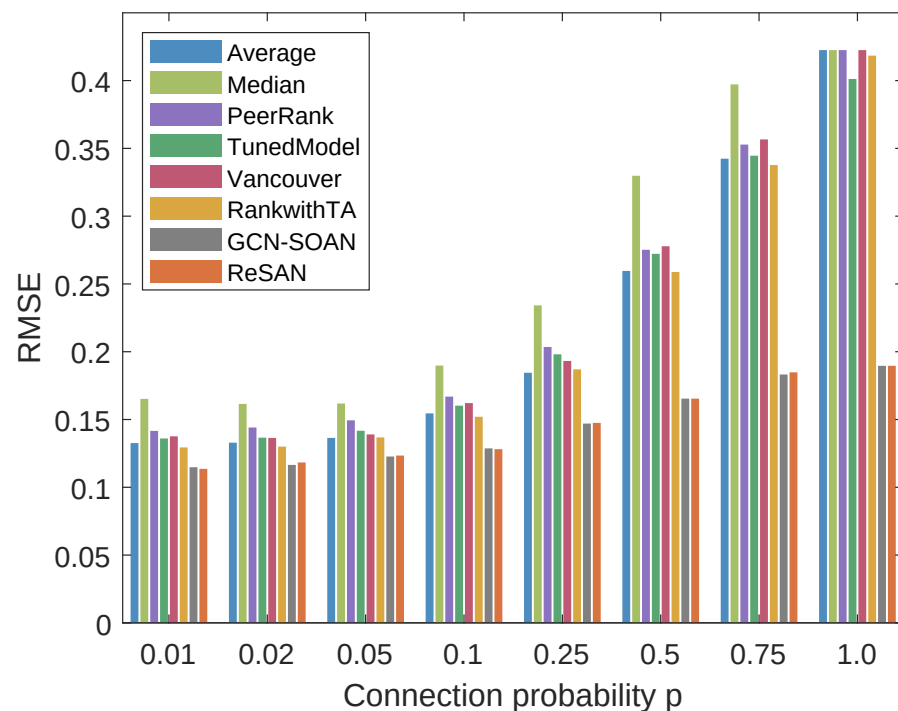


Figure 5. Mean RMSE from synthetic data using strategic peer grading and random graph models.

5.5. Parameter Sensitivity Analysis

In the ReSAN model, heads denotes the number of parallel attention layers in the self-attention mechanism. Multi-head attention allows the model to capture information from multiple representation subspaces, supporting a more comprehensive understanding of the relationships between evaluators and assessed items.

We evaluated the effect of heads on performance using a real-world peer evaluation dataset comprising four assignments, testing values from 1 to 6. As shown in Figure 6, the model's RMSE remained generally stable across different heads settings. While the lowest RMSE was observed with heads = 1, using multiple attention heads did not lead to substantial degradation, indicating that the model is robust to this parameter choice.

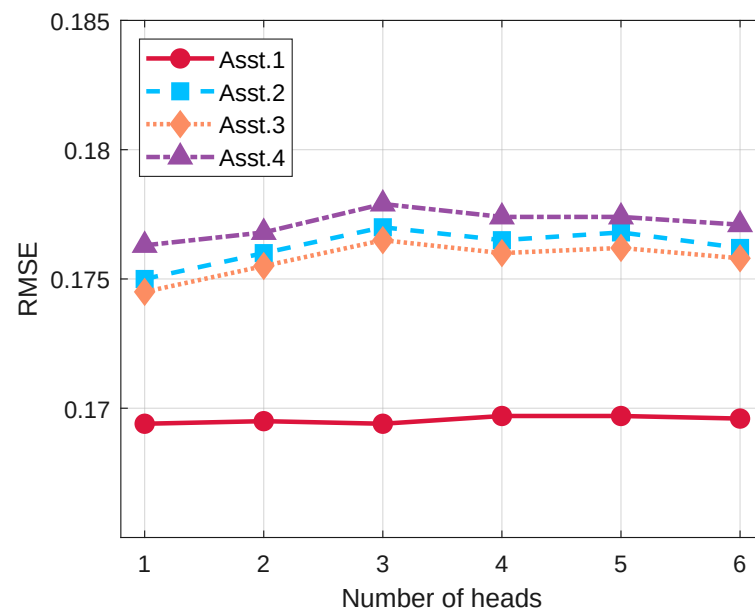


Figure 6. Parameter sensitivity analysis of multi-head attention mechanisms on the real-world datasets.

These findings suggest that ReSAN can effectively integrate multi-head attention to capture diverse aspects of scoring behavior, while maintaining stable predictive performance across a range of heads values.

5.6. Ablation Study

To further examine the role of edge weight in ReSAN, we conducted an ablation study by removing this component from the model. The motivation behind this experiment is to verify whether explicitly modeling the strength and reliability of evaluation relations contributes to prediction accuracy, or whether the model can perform comparably by treating all evaluation signals equally. For robustness, the study was carried out on both the real-world peer grading dataset and a controlled synthetic dataset.

The results are summarized in Tables 5 and 6. In the real-world dataset, the full model consistently outperforms the variant without edge weight across all four assignments. For example, under the Peer evaluation setting, the average error of the full model remains around 0.16–0.18, whereas the model without edge weight shows markedly higher errors, ranging from 0.24 to 0.32. A similar trend is observed under the Peer and self evaluation setting, where the absence of edge weight leads to substantial performance degradation.

Table 5. Ablation study on the contributions of edge_weight on the real_world datasets.

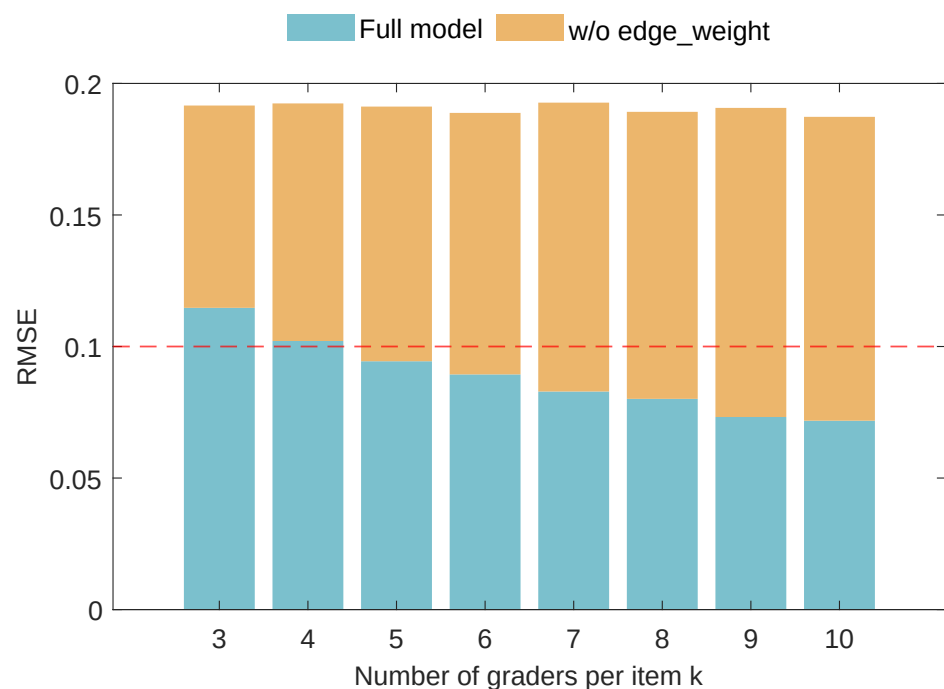
Model	Peer Evaluation			
	Asst. 1	Asst. 2	Asst. 3	Asst. 4
Full model	0.1690	0.1617	0.1835	0.1749
w/o edge_weight	0.2670	0.2444	0.3246	0.2746
Model	Peer and Self Evaluation			
	Asst. 1	Asst. 2	Asst. 3	Asst. 4
Full model	0.1668	0.1556	0.1829	0.1747
w/o edge_weight	0.2670	0.2444	0.3246	0.2746

Table 6. Ablation study on the contributions of edge_weight on the synthetic datasets.

Model	Number of Graders per Item k			
	3	4	5	6
Full model	0.1147	0.1021	0.0944	0.0894
w/o edge_weight	0.1916	0.1924	0.1912	0.1888

Model	Number of Graders per Item k			
	7	8	9	10
Full model	0.0829	0.0801	0.0732	0.0718
w/o edge_weight	0.1927	0.1892	0.1907	0.1873

On the synthetic dataset, we further examined the effect of varying the number of graders per item. The full model achieves steadily decreasing errors as the number of graders increases, dropping to around 0.07 when each item receives ten graders. By contrast, the performance of the model without edge weight remains relatively flat and significantly higher, fluctuating between 0.18 and 0.19 regardless of the number of graders. This indicates that, without edge weights, the model cannot fully benefit from additional grading signals. Through the visualization of the ablation study in Figure 7, the differences in model performance become even clearer.

**Figure 7.** Visualization of the ablation study results on the synthetic dataset.

In the ablation results, an interesting observation emerges: once edge weights are removed, the predictions under Peer evaluation and Peer and self evaluation become almost identical across all assignments, with no discernible difference under varying parameters. This phenomenon suggests that, in the absence of edge-weight guidance, all evaluation relations are treated uniformly during graph convolution, eliminating the distinction between peer and self-assessment signals. In other words, the model loses its ability to discriminate between sources of evaluation and falls back to relying solely on the structural connections of the graph.

In summary, these findings confirm the critical role of edge weight in ReSAN modeling. Edge weight not only encodes the diversity of assessment relations but also enables the model to differentiate signals from diverse sources, which is crucial for accurate score prediction.

6. Further Discussion

Our results confirm that adaptive relation-sensitive modeling can effectively improve the reliability of peer assessment by emphasizing trustworthy interactions and mitigating biased or inconsistent evaluations. This approach addresses key limitations of uniform aggregation methods and demonstrates strong robustness across diverse datasets.

From a practical perspective, the framework shows promise in supporting fair and transparent peer assessment in large-scale educational settings. By lessening instructors' grading burden and facilitating reliable peer feedback, it contributes to both operational efficiency and enhanced learner engagement. Compared to prior works, our method's ability to adapt relational weights contextually improves both model interpretability and predictive accuracy.

However, several limitations should be noted. The construction of the graph, particularly how relation types and weights are defined—can significantly impact performance. Moreover, the attention mechanism adds computational overhead, which may pose scalability challenges in very large cohorts. Another important limitation is transferability: since relational patterns and graph structures vary across courses or assignments, the model requires retraining for each new context. While this allows for flexible adaptation, it restricts direct generalization. To overcome these challenges and further improve the framework, we propose three promising directions for future research. First, exploring transfer learning or meta-learning strategies could enable the model to generalize relation-sensitive representations across different educational environments. Second, incorporating temporal dynamics would allow the model to capture how evaluator reliability evolves over time and across tasks. Third, integrating multimodal data (such as textual feedback and interaction logs) could enrich relational modeling with content-level information. These enhancements are expected to strengthen both interpretability and practical applicability of peer evaluation systems.

7. Conclusions

This study presented an adaptive relation-sensitive framework (ReSAN) designed to enhance the reliability of peer assessment. By dynamically weighting interactions within educational graphs, the method effectively distinguishes trustworthy evaluations from biased or inconsistent ones, overcoming the limitations of traditional uniform aggregation strategies. Experiments on both synthetic and real-world datasets demonstrate the robustness and generalizability of the proposed design. Methodologically, the framework introduces a novel relation-sensitive message passing mechanism that enables context-aware differentiation among evaluators.

Practically, ReSAN supports more credible and scalable peer assessment by reducing instructor workload while promoting fairness and engagement in large-scale learning environments. Looking forward, we plan to extend this framework by incorporating temporal graphs to capture evolving evaluation behaviors, integrating multimodal data such as textual comments or behavioral traces, and enhancing interpretability to provide actionable feedback. These future directions aim to further advance reliable, transparent, and effective peer evaluation in educational practice.

Author Contributions: Conceptualization, X.M. and S.Y.; Methodology, X.M., Y.F. and S.Y.; Software, Y.F., S.Z. and S.Y.; Validation, Y.F., Y.G., S.Z. and S.Y.; Formal analysis, Y.F.; Investigation, X.M.; Resources, X.M.; Data curation, X.M.; Writing—original draft, X.M., Y.F., Y.G., S.Z. and S.Y.; Writing—review & editing, X.M., Y.F., Y.G., S.Z. and S.Y.; Visualization, X.M., Y.F., Y.G., S.Z. and S.Y. All authors have read and agreed to the published version of the manuscript.

Funding: Shasha Yang acknowledged A Project Supported by Scientific Research Fund of Zhejiang Provincial Education Department (No. Y202353710).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original data presented in the study are openly available at https://www.tml.cs.uni-tuebingen.de/team/luxburg/code_and_data/peer_grading_data_request_new.php (accessed on 16 January 2025).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

For clarity, all abbreviations and mathematical notations introduced above are summarized here. The table provides a comprehensive reference for symbols used throughout the paper, covering graph fundamentals, the construction of the Social–Ownership–Assessment Network (SOAN), relation-sensitive attention mechanisms, and the learning objective. This summary is intended to improve readability and ensure consistency in notation across theoretical analysis and experimental sections:

Symbol	Definition
GNN	Graph Neural Network
ReSAN	Relation-Sensitive Assessment Network
SOAN	Social–Ownership–Assessment Network
RMSE	Root Mean Square Error
$\mathcal{G} = (\mathcal{V}, \mathcal{E})$	Graph with vertex set $\mathcal{V}$ and edge set $\mathcal{E}$
$N = \mathcal{V} $	Number of vertices
$M = \mathcal{E} $	Number of edges
$A \in \{0, 1\}^{N \times N}$	Adjacency matrix of the graph
$A(u, v)$	Entry of A ; $A(u, v) = 1$
$D \in \mathbb{R}^{N \times N}$	Degree matrix with $D(v) = \sum_{u \in \mathcal{V}} A(u, v)$
$\hat{A} = A + I$	Adjacency matrix with self-loops
$\hat{D}$	Corresponding degree matrix of $\hat{A}$
$X \in \mathbb{R}^{N \times d}$	Node feature matrix
$X^{(\ell)} \in \mathbb{R}^{N \times C}$	Node representation at layer ℓ
$\Theta^{(\ell)}$	Trainable weight matrix at layer ℓ
$\sigma(\cdot)$	Nonlinear activation function (e.g., ReLU or ELU)
$\mathcal{U}$	Set of users
$\mathcal{I}$	Set of items
$n = \mathcal{U} $	Number of users
$m = \mathcal{I} $	Number of items
v_i	True (ground-truth) score of item i
$\hat{v}_i$	Predicted score of item i
$A = [a_{ui}] \in \mathbb{R}^{n \times m}$	Assessment matrix; a_{ui} is score given by user u to item i
$S \in \mathbb{R}^{n \times n}$	Social adjacency matrix among users
$O \in \mathbb{R}^{n \times m}$	Ownership matrix between users and items
$N = O + A$	Combined user–item relation matrix
M	Block matrix encoding SOAN structure
$\mathcal{N}(i)$	Neighborhood of node i defined by M

$h_i^{(\ell)} \in \mathbb{R}^d$	Embedding of node i at layer ℓ
$W^{(\ell)} \in \mathbb{R}^{d' \times 2d}$	Linear projection matrix for attention computation
$a \in \mathbb{R}^{d'}$	Attention vector
$e_{ij}^{(\ell)}$	Raw (unnormalized) attention score between nodes i and j
$\alpha_{ij}^{(\ell)}$	Normalized attention coefficient via softmax
L	Total number of layers in the model
w, b	Parameters of the regression head for score prediction
$\mathcal{L}$	Loss function (Root Mean Squared Error, RMSE)
$\mathcal{D}$	Set of items with known ground-truth scores

References

1. Topping, K.J. Peer assessment. *Theory Pract.* **2009**, *48*, 20–27. [\[CrossRef\]](#)
2. Formanek, M.; Wenger, M.C.; Buxner, S.R.; Impey, C.D.; Sonam, T. Insights about large-scale online peer assessment from an analysis of an astronomy MOOC. *Comput. Educ.* **2017**, *113*, 243–262. [\[CrossRef\]](#)
3. Alcarria, R.; Bordel, B.; De Andra, D.M. Enhanced peer assessment in MOOC evaluation through assignment and review analysis. *Int. J. Emerg. Technol. Learn. (ijET)* **2018**, *13*, 206–219. [\[CrossRef\]](#)
4. Berkman, F.; Bigerelle, M.; Lemesle, J.; Nys, L.; Wiczorowski, M.; Brown, C. Peer Assessment in Interdisciplinary Learning: Measuring Reliability and Engaging Critical Thinking. *Think. Ski. Creat.* **2025**, *58*, 101950. [\[CrossRef\]](#)
5. Seifert, T.; Feliks, O. Online self-assessment and peer-assessment as a tool to enhance student-teachers' assessment skills. *Assess. Eval. High. Educ.* **2019**, *44*, 169–185. [\[CrossRef\]](#)
6. Garcia-Loro, F.; Martin, S.; Ruipérez-Valiente, J.A.; Sancristobal, E.; Castro, M. Reviewing and analyzing peer review Inter-Rater Reliability in a MOOC platform. *Comput. Educ.* **2020**, *154*, 103894. [\[CrossRef\]](#)
7. Perdue, M.; Sandland, J.; Joshi, A.; Liu, J. Exploring the Integration of Social Practice into MOOC Peer Assessment. In Proceedings of the 2024 IEEE Digital Education and MOOCs Conference (DEMOcon), Atlanta, GA, USA, 16–18 October 2024; pp. 1–6.
8. Topping, K.J.; Gehringer, E.; Khosravi, H.; Gudipati, S.; Jadhav, K.; Susarla, S. Enhancing peer assessment with artificial intelligence. *Int. J. Educ. Technol. High. Educ.* **2025**, *22*, 3. [\[CrossRef\]](#)
9. Sajjadi, M.S.M.; Alamgir, M.; von Luxburg, U. Peer grading in a course on algorithms and data structures: Machine learning algorithms do not improve over simple baselines. In Proceedings of the Third ACM Conference on Learning@ Scale, Scotland, UK, 25–26 April 2016; pp. 369–378.
10. Piech, C.; Huang, J.; Chen, Z.; Do, C.B.; Ng, A.; Koller, D. Tuned Models of Peer Assessment in MOOCs. In Proceedings of the 6th International Conference on Educational Data Mining, Memphis, TN, USA, 6–9 July 2013.
11. Walsh, T. The PeerRank method for peer assessment. In Proceedings of the Twenty-First European Conference on Artificial Intelligence, Prague, Czech Republic, 18–22 August 2014; pp. 909–914.
12. de Alfaro, L.; Shavlovsky, M. CrowdGrader: A tool for crowdsourcing the evaluation of homework assignments. In Proceedings of the 45th ACM Technical Symposium on Computer Science Education, Atlanta, GA, USA, 5–8 March 2014; pp. 415–420.
13. Fang, H.; Wang, Y.; Jin, Q.; Ma, J. RankwithTA: A robust and accurate peer grading mechanism for MOOCs. In Proceedings of the 2017 IEEE 6th International Conference on Teaching, Assessment, and Learning for Engineering (TALE), Hong Kong, China, 12–14 December 2017; pp. 497–502.
14. Wang, T.; Jing, X.; Li, Q.; Gao, J.; Tang, J. Improving Peer Assessment Accuracy by Incorporating Relative Peer Grades. In Proceedings of the International Educational Data Mining Society, Montreal, QC, Canada, 2–5 July 2019.
15. Mubarak, A.A.; Cao, H.; Hezam, I. M.; Hao, F. Modeling students' performance using graph convolutional networks. *Complex Intell. Syst.* **2022**, *8*, 2183–2201. [\[CrossRef\]](#)
16. Dawson, S. A study of the relationship between student social networks and sense of community. *J. Educ. Technol. Soc.* **2008**, *11*, 224–238.
17. Williams, R.T. An overview of MOOCs and blended learning: Integrating MOOC technologies into traditional classes. *IETE J. Educ.* **2024**, *65*, 84–91. [\[CrossRef\]](#)
18. Papadakis, S. MOOCs 2012–2022: An overview. *Adv. Mob. Learn. Educ. Res.* **2023**, *3*, 682–693. [\[CrossRef\]](#)
19. Tzeng, J.-W.; Lee, C.-A.; Huang, N.-F.; Huang, H.-H.; Lai, C.-F. MOOC evaluation system based on deep learning. *Int. Rev. Res. Open Distrib. Learn.* **2022**, *23*, 21–40. [\[CrossRef\]](#)
20. Gamage, D.; Staubitz, T.; Whiting, M. Peer assessment in MOOCs: Systematic literature review. *Distance Educ.* **2021**, *42*, 268–289. [\[CrossRef\]](#)
21. Ortega-Ruipérez, B.; Correa-Gorospe, J.M. Peer assessment to promote self-regulated learning with technology in higher education: Systematic review for improving course design. *Front. Educ.* **2024**, *9*, 1376505. [\[CrossRef\]](#)

22. Paul, U.; Mantravadi, A.; Shah, J.; Shah, S.; Mylavarapu, S. V.; Rashid, M. P.; Gehringer, E. Scaling Success: A Systematic Review of Peer Grading Strategies for Accuracy, Efficiency, and Learning in Contemporary Education. *arXiv* **2025**, arXiv:2508.11677.
23. Xiong, Y.; Schunn, C. D.; Wu, Y. What predicts variation in reliability and validity of online peer assessment? A large-scale cross-context study. *J. Comput. Assist. Learn.* **2023**, *39*, 2004–2024. [[CrossRef](#)]
24. Morris, W.; Crossley, S.; Holmes, L.; Trumbore, A. Using transformer language models to validate peer-assigned essay scores in massive open online courses (MOOCs). In Proceedings of the 13th International Learning Analytics and Knowledge Conference, Arlington, TX, USA, 13–17 March 2023; pp. 315–323.
25. Kipf, T.N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. In Proceedings of the 5th International Conference on Learning Representations, Toulon, France, 24–26 April 2017.
26. Velickovic, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; Bengio, Y. Graph attention networks. *Stat* **2017**, *1050*, 10–48550.
27. Hamilton, W.; Ying, Z.; Leskovec, J. Inductive representation learning on large graphs. In Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
28. Hansen, J.; Gebhart, T. Sheaf Neural Networks. In Proceedings of the TDA & Beyond Workshop, Vancouver, BC, Canada, 11 December 2020.
29. Zheng, X.; Zhou, B.; Gao, J.; Wang, Y. G.; Li'o, P.; Li, M.; Mont'ufar, G. How framelets enhance graph neural networks. In Proceedings of the 38th International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 12761–12771.
30. Sharma, K.; Lee, Y.-C.; Nambi, S.; Salian, A.; Shah, S.; Kim, S.-W.; Kumar, S. A survey of graph neural networks for social recommender systems. *ACM Comput. Surv.* **2024**, *56*, 1–34. [[CrossRef](#)]
31. Guo, Z.; Wang, H. A deep graph neural network-based mechanism for social recommendations. *IEEE Trans. Ind. Inform.* **2020**, *17*, 2776–2783. [[CrossRef](#)]
32. Liu, T.; Wang, Y.; Ying, R.; Zhao, H. Muse-gnn: Learning unified gene representation from multimodal biological graph data. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 24661–24677.
33. Li, S.; Hua, H.; Chen, S. Graph neural networks for single-cell omics data: A review of approaches and applications. *Briefings Bioinform.* **2025**, *26*, bbaf109. [[CrossRef](#)]
34. Namanloo, A.A.; Thorpe, J.; Salehi-Abari, A. Improving Peer Assessment with Graph Neural Networks. In Proceedings of the International Educational Data Mining Society, Durham, UK, 24–27 July 2022.
35. Li, M.; Cheng, Y.; Bai, L.; Cao, F.; Lv, K.; Liang, J.; Lio, P. EduLLM: Leveraging Large Language Models and Framelet-Based Signed Hypergraph Neural Networks for Student Performance Prediction. In Proceedings of the 42nd International Conference on Machine Learning, Vancouver, BC, Canada, 13–19 July 2025.
36. Li, M.; Zhou, S.; Chen, Y.; Huang, C.; Jiang, Y. EduCross: Dual adversarial bipartite hypergraph learning for cross-modal retrieval in multimodal educational slides. *Inf. Fusion* **2024**, *109*, 102428. [[CrossRef](#)]
37. Schlichtkrull, M.; Kipf, T. N.; Bloem, P.; Van Den Berg, R.; Titov, I.; Welling, M. Modeling relational data with graph convolutional networks. In Proceedings of the 15th European Semantic Web Conference, Heraklion, Greece, 3–7 June 2018; pp. 593–607.
38. Vashishth, S.; Sanyal, S.; Nitin, V.; Talukdar, P. Composition-based Multi-Relational Graph Convolutional Networks. In Proceedings of the 8th International Conference on Learning Representations, Addis Ababa, Ethiopia, 26–30 April 2020.
39. Xu, K.; Hu, W.; Leskovec, J.; Jegelka, S. How Powerful are Graph Neural Networks? In Proceedings New Orleans, LA, USA, 6–9 May 2019.
40. Brody, S.; Alon, U.; Yahav, E. How Attentive are Graph Attention Networks? In Proceedings of the International Conference on Learning Representations, Virtual, 25–29 April 2022.
41. Wang, X.; Ji, H.; Shi, C.; Wang, B.; Ye, Y.; Cui, P.; Yu, P.S. Heterogeneous graph attention network. In Proceedings of the 2019 World Wide Web Conference, San Francisco, CA, USA, 13–17 May 2019; pp. 2022–2032.
42. Wang, Q.; Mao, Z.; Wang, B.; Guo, L. Knowledge graph embedding: A survey of approaches and applications. *IEEE Trans. Knowl. Data Eng.* **2017**, *29*, 2724–2743. [[CrossRef](#)]
43. Gao, C.; Zheng, Y.; Li, N.; Li, Y.; Qin, Y.; Piao, J.; Quan, Y.; Chang, J.; Jin, D.; He, X. A survey of graph neural networks for recommender systems: Challenges, methods, and directions. *ACM Trans. Recomm. Syst.* **2023**, *1*, 1–51. [[CrossRef](#)]
44. Ying, R.; He, R.; Chen, K.; Eksombatchai, P.; Hamilton, W.L.; Leskovec, J. Graph convolutional neural networks for web-scale recommender systems. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK, 19–23 August 2018; pp. 974–983.
45. Gilmer, J.; Schoenholz, S.S.; Riley, P.F.; Vinyals, O.; Dahl, G.E. Neural message passing for quantum chemistry. In Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 1263–1272.
46. Kulkarni, C.; Wei, K.P.; Le, H.; Chia, D.; Papadopoulos, K.; Cheng, J.; Koller, D.; Klemmer, S.R. Peer and self assessment in massive online classes. *ACM Trans.-Comput.-Hum. Interact.* **2013**, *20*, 1–31. [[CrossRef](#)]
47. Suen, H.K. Peer assessment for massive open online courses (MOOCs). *Int. Rev. Res. Open Distrib. Learn.* **2014**, *15*, 312–327. [[CrossRef](#)]

48. Alves, T.; Sousa, F.; Gama, S.; Jorge, J.; Gonçalves, D. How personality traits affect peer assessment in distance learning. *Technol. Knowl. Learn.* **2024**, *29*, 371–396. [\[CrossRef\]](#)
49. Yang, G.; Li, M.; Feng, H.; Zhuang, X. Deeper insights into deep graph convolutional networks: Stability and generalization. *IEEE Trans. Pattern Anal. Mach. Intell.* **2025**, *Early Access*. [\[CrossRef\]](#)
50. Xu, Q.S.; Liang, Y.Z. Monte Carlo cross validation. *Chemom. Intell. Lab. Syst.* **2001**, *56*, 1–11. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.