Latest updates: https://dl.acm.org/doi/10.1145/3690624.3709264

RESEARCH-ARTICLE

Exploring Heterogeneity and Uncertainty for Graph-based Cognitive Diagnosis Models in Intelligent Education

PENGYANG SHAO, Hefei University of Technology, Hefei, Anhui, China

YONGHUI YANG, Hefei University of Technology, Hefei, Anhui, China

CHEN GAO, Beijing National Research Center for Information Science and Technology, Beijing, China

LEI CHEN, Beijing National Research Center for Information Science and Technology, Beijing, China

KUN ZHANG, Hefei University of Technology, Hefei, Anhui, China

CHENYI ZHUANG, Ant group, Hangzhou, Zhejiang, China

View all

Open Access Support provided by:

Hefei University of Technology

Beijing National Research Center for Information Science and Technology

Ant group

Exploring Heterogeneity and Uncertainty for Graph-based Cognitive Diagnosis Models in Intelligent Education

Pengyang Shao*
Key Laboratory of Knowledge
Engineering with Big Data, Hefei
University of Technology
Hefei, Anhui, China

Yonghui Yang*
Key Laboratory of Knowledge
Engineering with Big Data, Hefei
University of Technology
Hefei, Anhui, China

Chen Gao
BNRist, Tsinghua University
Beijing, China

Lei Chen
Department of Electronic
Engineering, BNRist, Tsinghua
University
Beijing, China

Kun Zhang
Key Laboratory of Knowledge
Engineering with Big Data, Hefei
University of Technology
Hefei, Anhui, China

Chenyi Zhuang
Ant Group
Hangzhou, Zhejiang, China

Le Wu
Key Laboratory of Knowledge
Engineering with Big Data, Hefei
University of Technology
Hefei, Anhui, China

Yong Li
Department of Electronic
Engineering, BNRist, Tsinghua
University
Beijing, China

Meng Wang†
Key Laboratory of Knowledge
Engineering with Big Data, Hefei
University of Technology
Hefei, Anhui, China
eric.mengwang@gmail.com

Abstract

Graph-based Cognitive Diagnosis (CD) has attracted much research interest due to its strong ability on inferring students' proficiency levels on knowledge concepts. While graph-based CD models have demonstrated remarkable performance, we contend that they still cannot achieve optimal performance due to the neglect of edge heterogeneity and uncertainty. Edges involve both correct and incorrect response logs, indicating heterogeneity. Meanwhile, a response log can have uncertain semantic meanings, e.g., *a correct log can indicate true mastery or fortunate guessing, and a wrong log can indicate a lack of understanding or a careless mistake.* In this paper, we propose an **I**nformative **S**emantic-aware **G**raph-based **C**ognitive **D**iagnosis model (**ISG-CD**), which focuses on how to utilize the heterogeneous graph in CD and minimize effects of uncertain edges. Specifically, to explore heterogeneity, we propose a semantic-aware graph neural networks based CD model. To minimize effects of edge uncertainty, we propose an Informative Edge Differentiation layer from an information bottleneck perspective, which suggests keeping a minimal yet sufficient reliable graph for CD in an unsupervised way. We formulate this process as maximizing mutual information between the reliable graph and response logs, while minimizing mutual information between the reliable graph and the original graph. After that, we prove that mutual information maximization can be theoretically converted to the classic binary cross entropy loss function, while minimizing mutual information can be realized by the Hilbert-Schmidt Independence Criterion. Finally, we adopt an alternating training strategy for optimizing learnable parameters of both the semantic-aware graph neural networks based CD model and the edge differentiation layer. Extensive experiments on three real-world datasets have demonstrated the effectiveness of ISG-CD.

CCS Concepts

• **Applied computing** → **E-learning**.

Keywords

Cognitive Diagnosis, Graph Neural Network, Student Modeling

*Equal Contribution.
†Corresponding Author.

1 Introduction

Educational CD has attracted much research interest due to growing needs in online education [11, 36, 54]. Neural network-based CD models have garnered attention due to detailed diagnosis and relatively-high performance [19, 24, 37, 41, 50]. As shown in Figure 1 (a), these models take students' response logs, and expert-labeled exercise-concept relations as input, and output students' proficiency
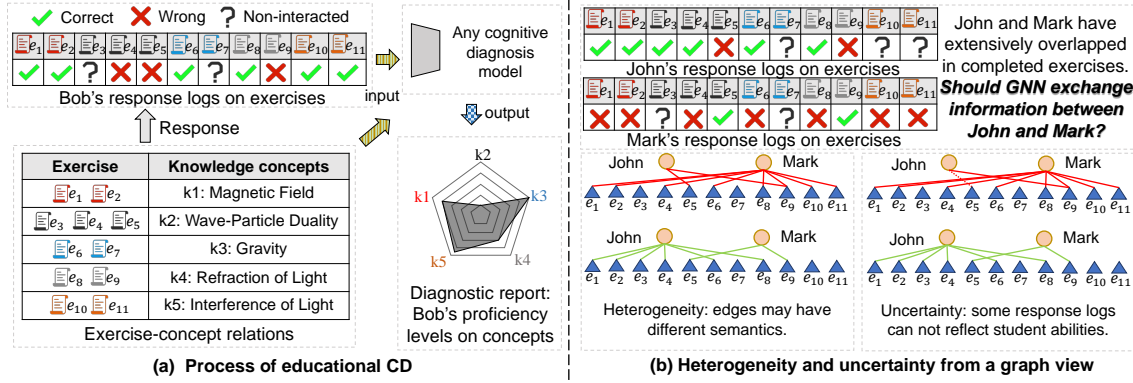
Figure 1: (a) The process of educational CD. These CD models take students' responses logs on exercises, and exercise-concept relations as input, and output students' proficiency levels on all knowledge concepts. (b) GNN between two students (John and Mark). Their interactions with exercises have extensively overlapped, however, they are not similar from a graph view due to heterogeneity. Further, there are uncertain edges, e.g., according to interactions between John and exercises e_3, e_4, John has already grasped concept $k2$. John's performance on exercise e_5 cannot reflect his ability on concept $k2$.

levels on all concepts. The cornerstone lies in a diagnostic function that establishes connections between all trainable parameters and student-exercise response logs [39].

Researchers have found that incorporating Graph Neural Network (GNN) into CD models can further improve diagnostic performance [14, 42]. Typically, response logs naturally form a student-exercise bipartite graph. Each node in the graph represents either a student or an exercise, and each edge represents a student-exercise response log. As GNNs have powerful feature learning and pattern extraction capabilities [4], a natural idea is to adopt GNNs in capturing high-order graph information, thus enhancing the quality of student and exercise representations. These enhanced representations lead to more accurate predictions of students' proficiency levels on concepts.

While graph-based CD models have demonstrated remarkable performance, we contend that they still have inherent limitations [14, 42]. On one hand, edges may correspond to both correct and incorrect response logs, indicating edge **heterogeneity**. As shown in Figure 1 (b), we consider: *"Should GNN exchange information between John's and Mark's representations?"* As John and Mark have extensively overlapped in completed exercises, previous studies may exchange information between them [14, 42]. However, they provide different answers to each exercise, indicating that their abilities are fundamentally opposite. Hence, we argue that GNN should not propagate information between them. On the other hand, each response log may have **uncertainty** in semantics, i.e., a correct log can indicate true mastery or fortunate guessing, and a wrong log can indicate a lack of understanding or a careless mistake. As shown in Figure 1 (b), John has correctly answered e_3, e_4 related to concept $k2$ but wrongly answered e_5 related to the same concept. It is hard to tell whether John's response to e_5 is due to a lack of ability or a careless mistake. We want to ensure that such uncertain response logs would not affect the process of GNN. Due to lack of labels, how to alleviate effects of uncertainty in an unsupervised way becomes our primary focus.

In this paper, we propose a novel **I**nformative **S**emantic-aware **G**raph-based **C**ognitive **D**iagnosis model (**ISG-CD**), which focuses on how to utilize the heterogeneous graph in CD and minimize effects of uncertain edges. Specifically, we take KaNCD as our backbone model due to its relatively high performance and easy-to-implementation [39]. KaNCD utilizes an embedding module that maps students, exercises, and concepts to a latent representation space. Then, the matrix factorization layer takes students' and concepts' latent representations as input, and outputs students' proficiency levels on concepts. Exercise difficulties can be obtained analogously. Compared to KaNCD, our improvements can be reflected in two aspects. First, to capture edge heterogeneity, we propose a **Semantic-aware GNN (S-GNN)** based CD model. Inspired by previous works [4], we first obtain two subgraphs based on edge semantics (correct/incorrect response logs), and then update student and exercise embeddings based on these subgraphs. Second, to minimize effects of uncertain edges, we propose an **Informative Edge Differentiation (IE-Diff)** layer which can obtain a reliable student-exercise graph. Due to the lack of uncertainty labels, IE-Diff layer is designed based on Information Bottleneck (IB) principle [29, 34], which can keep a minimal yet sufficient reliable graph for CD in an unsupervised way. We parameterize the reliable graph based on the probability of dropping edges. Obviously, the reliable graph should assign high probabilities to those uncertain edges, and assign low probabilities to certain edges. To model the probability of dropping edges, we use Bernoulli distributions, whose parameters are estimated from corresponding student and exercise latent embeddings. Then, we summarize the goal of IE-Diff as maximizing mutual information between the reliable graph and response logs, while minimizing mutual information between the reliable graph and the original graph. Directly minimizing/maximizing mutual information is difficult, therefore, we prove that mutual information maximization between the reliable graph and response logs can be theoretically converted to the classic binary cross entropy loss function, while minimizing mutual information between the reliable graph and the original graph can

be realized by the Hilbert-Schmidt Independence Criterion [16, 25]. Finally, we adopt an alternating training strategy to together optimize learnable parameters of the S-GNN based CD model and the IE-Diff layer [58, 59]. Extensive experiments on three real-world datasets have demonstrated the effectiveness of ISG-CD. The major contributions can be summarized as follows:

- We explore semantic heterogeneity and uncertainty in the student-exercise bipartite graph. We argue that neglecting heterogeneity and uncertainty will lead to sub-optimal diagnostic performance of graph-based CD models.
- We propose a novel ISG-CD model, which focuses on how to utilize the heterogeneous graph and minimize effects of uncertain edges. We first propose a semantic-aware GNN layer to utilize the graph, and then propose an informative edge differentiation layer to handle uncertainty.
- Extensive experiments on three real-world datasets have demonstrated the stable effectiveness of ISG-CD, e.g., ISG-CD achieves stable improvements of over 1% than best baselines on the DOA metric on all three datasets.

2 Related Work

2.1 Educational Cognitive Diagnosis

Research on educational CD has primarily emerged from the field of psychology. Two classic methods in this area are Item Response Theory (IRT) [11] and DINA [10]. IRT utilizes continuous one dimensional parameters to represent student and exercise entities. Further, it employs a logistic function to infer students' performance on exercises. Multidimensional Item Response Theory (MIRT) has been proposed for better ability estimations by extending parameter dimensions [28]. DINA represents student and exercise entities with binary discrete variables. Then, it introduces "slip" and "guess" parameters to better fit real-world scenarios [10]. With the recent advancements in neural networks, researchers have turned their attention to leveraging neural networks in educational CD [9, 15, 20, 50, 57]. NCDM framework first utilizes high-dimensional continuous representations to model students' abilities and exercise difficulties and proposes a novel diagnostic function to incorporate all parameters to obtain final results [38]. Inspired by collaborative filtering [6, 8, 56], KaNCD utilizes a matrix factorization layer to fuse students' and concepts' latent representations for students' proficiency levels on concepts [39].

Building upon the success of the NCDM framework, graph-based CD models integrate graph structure information into CD models [23, 27, 31]. RCD simultaneously captures graph information hidden in the student-exercise bipartite graph and concept relationships [14]. Compared to RCD, SCD introduces a graph self-supervised learning to boost student prediction performance [42]. ORCDF further proposes a response-aware GNN layer with a graph contrastive learning based regularization term to enhance the robustness of graph-based CD models [27]. Compared to these graph-based CD models, our distinction lies in exploring the semantic edge heterogeneity and minimizing effects of uncertain edges.

2.2 Uncertainty Detection in GNN

Researchers point out that there may exist uncertainty in a graph structure, leading to suboptimal performance of GNN. An intuitive approach is to adaptively select node features or edges [7, 21, 55], e.g., updating node features [48], modifying edges [51]. Another approach focuses on enhancing the robustness of GNN to the input graph structure, e.g., self-supervised learning based methods [14], or adversarial learning based methods [32].

Among these approaches, graph refinements based on information bottleneck principles have garnered significant attention due to relatively-high performance and efficiency [33, 49]. These principles mainly try to capture minimal yet sufficient information hidden in the graph structure while discarding information irrelevant to downstream tasks [49]. Researchers have substantiated that these principles can help GNN effectively mitigate adversarial attacks [40]. Yang et al. successfully preserve minimal yet sufficient information in social graphs for social recommendations [52]. In this paper, we propose to minimize effects of edge uncertainty from the information bottleneck perspective. This process requires maximizing mutual information between a reliable graph and response logs, and minimizing mutual information between the reliable graph and the original graph.

3 The Proposed Model

In this section, we first present important notations and task formulation, followed by the overall structure of our proposed ISG-CD. After that, we introduce each component of ISG-CD in detail, and alternating training for ISG-CD. Finally, we discuss the time and space complexity of ISG-CD.

3.1 Task Formulation

There are students $S(|S| = M)$, exercises $E(|E| = N)$, and concepts $K(|K| = T)$. There are two types of relationships. First, students practice some exercises, forming response logs. r_{se} is student s's response log to exercise e. We use $R = \{(s, e, r_{se})\}$ to denote triplets of students, exercises and response logs. If student s answers exercise e correctly, $r_{se} = 1$. Otherwise, $r_{se} = 0$. In this paper, we also treat response logs as a bipartite graph $\mathcal{G}$. Second, the relations between exercises and concepts are denoted by $Q = \{q_{ek}\}_{N \times T}$. If exercise e is related to concept k, $q_{ek} = 1$; otherwise, $q_{ek} = 0$. Usually, Q is pre-defined by experts. $A = [a_1, ..., a_s, ..., a_M]^\top \in \mathbb{R}^{M \times T}$ represents student abilities (i.e., proficiency levels on all concepts), where a_s denotes student s's ability. Each dimension of $a_s = [a_{s1}, ..., a_{sk}, ..., a_{sT}]$ has independent meaning, e.g., a_{sk} denotes student s's proficiency level on concept k. There are two exercise-side parameters, i.e., exercise difficulties $D = [d_1, ..., d_e, ..., d_N]^\top \in \mathbb{R}^{N \times T}$ and discriminations $h^{disc} = [h_1^{disc}, ..., h_e^{disc}, ..., h_N^{disc}]^\top \in \mathbb{R}^{N \times 1}$. Based on these notations, we formulate CD as follows,

Input: The response logs R and exercise-concept relations Q.
Output: A model to diagnose students' abilities (proficiency levels on all concepts) through response log prediction.

3.2 Overall Structure of ISG-CD

Compared to the backbone KaNCD [39], there are two additional components in ISG-CD. Figure 2 describes a semantic-aware GNN layer based CD model. S-GNN focuses on how to make full use of a reliable student-exercise graph for CD. After that, Figure 3 describes the informative edge differentiation (IE-Diff) layer, which
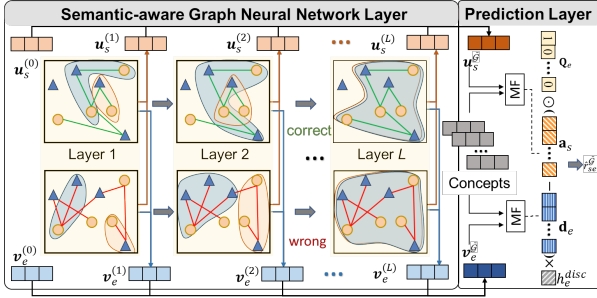
Figure 2: Semantic-aware GNN (S-GNN) based CD model.

Figure 3: Informative Edge Differentiation (IE-Diff) Layer.

is designed for obtaining the reliable graph based on information bottleneck principles [45].

These two layers are complementary to each other. On one hand, the IE-Diff layer relies on student and exercise latent embeddings from the S-GNN layer. On the other hand, the S-GNN layer relies on the reliable graph provided by the IE-Diff layer. Hence, we adopt alternately training these two layers [58, 59]. In the following parts of this section, we will discuss these two components.

3.3 Semantic-aware GNN Based CD Model

3.3.1 Embedding Module and Semantic-aware GNN (S-GNN) Layer.
Figure 2 describes incorporating the S-GNN layer into a CD model. We adopt KaNCD as our backbone due to its relatively-high performance and easy-to-implementation [39]. There are three latent embeddings in KaNCD, i.e., $\mathbf{U} = [\mathbf{u}_1, ..., \mathbf{u}_s, ..., \mathbf{u}_M]^\top \in \mathbb{R}^{M \times Z}$ for students, $\mathbf{V} = [\mathbf{v}_1, ..., \mathbf{v}_e, ..., \mathbf{v}_N]^\top \in \mathbb{R}^{N \times Z}$ for exercises and $\mathbf{O} = [\mathbf{o}_1, ..., \mathbf{o}_k, ..., \mathbf{o}_T]^\top \in \mathbb{R}^{T \times Z}$ for knowledge concepts. Z denotes the dimension of latent embeddings.

Inspired by recent success on graph learning [14, 17, 19], we focus on utilizing the student-exercise graph to enhance student and exercise latent embeddings. There are two types of edges with fundamentally different semantic meanings in this graph: correct answers and incorrect answers. Previous studies ignored this edge heterogeneity, which poses a risk of confounding semantic information in GNN-based CD models [14, 42]. To capture heterogeneity, we propose a S-GNN layer. Here, we suppose that we have already obtained the reliable graph $\tilde{\mathcal{G}}$. We split $\tilde{\mathcal{G}}$ into subgraphs based on semantics ($\tilde{\mathcal{G}} = \tilde{\mathcal{G}}_1 \cup \tilde{\mathcal{G}}_0$). $\tilde{\mathcal{G}}_1, \tilde{\mathcal{G}}_0$ denote subgraphs corresponding to correct/incorrect response logs, respectively. We treat latent embeddings of students and exercises in KaNCD as local embeddings, e.g., $\mathbf{v}_e = \mathbf{v}_e^{(0)}$, $\mathbf{u}_s = \mathbf{u}_s^{(0)}$. Edge-specific messages μ from exercise e to student s can be divided into two parts:

$$\mu_{e \to s, \tilde{\mathcal{G}}_1}^{(l)} = \frac{1}{\sqrt{|\mathcal{N}_{s,\tilde{\mathcal{G}}_1}||\mathcal{N}_{e,\tilde{\mathcal{G}}_1}|}}\mathbf{v}_e^{(l-1)}, \mu_{e \to s, \tilde{\mathcal{G}}_0}^{(l)} = \frac{1}{\sqrt{|\mathcal{N}_{s,\tilde{\mathcal{G}}_0}||\mathcal{N}_{e,\tilde{\mathcal{G}}_0}|}}\mathbf{v}_e^{(l-1)},$$ (1)

where $\mu_{e \to s, \tilde{\mathcal{G}}_1}^{(l)}$ and $\mu_{e \to s, \tilde{\mathcal{G}}_0}^{(l)}$ denote message passed in subgraph $\tilde{\mathcal{G}}_1$ and $\tilde{\mathcal{G}}_0$ at layer l, respectively. $\mathcal{N}_{s,\tilde{\mathcal{G}}_1}, \mathcal{N}_{e,\tilde{\mathcal{G}}_1}$ denote the neighbor sets of student s and exercise e in subgraph $\tilde{\mathcal{G}}_1$. Similarly, $\mathcal{N}_{s,\tilde{\mathcal{G}}_0}, \mathcal{N}_{e,\tilde{\mathcal{G}}_0}$ denote the neighbor sets in subgraph $\tilde{\mathcal{G}}_0$. $\mathbf{v}_e^{(l-1)}$ denotes exercise e's difficulty at layer $(l-1)$. Messages from student s to exercise e can
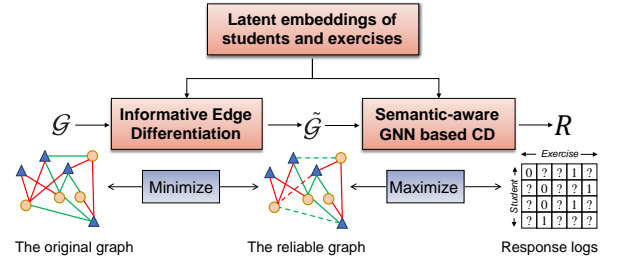
be processed analogously, formulated as:

$$\mu_{s \to e, \tilde{\mathcal{G}}_1}^{(l)} = \frac{1}{\sqrt{|\mathcal{N}_{s,\tilde{\mathcal{G}}_1}||\mathcal{N}_{e,\tilde{\mathcal{G}}_1}|}}\mathbf{u}_s^{(l-1)}, \mu_{s \to e, \tilde{\mathcal{G}}_0}^{(l)} = \frac{1}{\sqrt{|\mathcal{N}_{s,\tilde{\mathcal{G}}_0}||\mathcal{N}_{e,\tilde{\mathcal{G}}_0}|}}\mathbf{u}_s^{(l-1)},$$ (2)

where $\mathbf{u}_s^{(l-1)}$ denotes student s's latent embedding at layer $(l-1)$. After message passing, we can obtain incoming messages at every node from two types of edges. In message aggregation, for each node, we first sum incoming messages over all neighbors under a specific edge type. Subsequently, we accumulate the messages from all edge-types into vector representations at the l-th layer:

$$\mathbf{u}_s^{(l)} = \sum_e \mu_{e \to s, \tilde{\mathcal{G}}_1} + \sum_e \mu_{e \to s, \tilde{\mathcal{G}}_0}, \ \mathbf{v}_e^{(l)} = \sum_s \mu_{s \to e, \tilde{\mathcal{G}}_1} + \sum_s \mu_{s \to e, \tilde{\mathcal{G}}_0}.$$ (3)

We consider the following read out functions for final embeddings: $\mathbf{u}_s^{\tilde{\mathcal{G}}} = \sum_{l=0}^L \mathbf{u}_s^{(l)}, \mathbf{v}_e^{\tilde{\mathcal{G}}} = \sum_{l=0}^L \mathbf{v}_e^{(l)}$. The superscripts of $\mathbf{u}_s^{\tilde{\mathcal{G}}}$ and $\mathbf{v}_e^{\tilde{\mathcal{G}}}$ indicate that these representations are derived from graph $\tilde{\mathcal{G}}$.

3.3.2 Matrix Factorization and Prediction Layer. As shown in Figure 2, we combine latent embeddings to obtain students' proficiency levels and exercise difficulties via matrix factorization following KaNCD [39]. In this paper, we adopt a simple yet effective probabilistic matrix factorization, which is widely adopted in recommender systems [13, 26, 46]. The process is as follows:

$$a_{sk} = \sigma(<\mathbf{u}_s^{\tilde{\mathcal{G}}}, \mathbf{o}_k>), \ d_{ek} = \sigma(<\mathbf{v}_e^{\tilde{\mathcal{G}}}, \mathbf{o}_k>),$$ (4)

where $<,>$ denotes the inner dot between two embeddings. σ denotes the sigmoid activation function. By repeating Eq.(4) on all concepts, we can obtain student's proficiency levels and exercise difficulties, i.e., $\mathbf{a}_s$ and $\mathbf{d}_e$. Following [30], we allow adding bias terms during the above process. The connections between $\mathbf{a}_s, \mathbf{d}_e$ and response logs can be represented as that a student is more likely to answer an exercise correctly when the student's ability exceeds the exercise difficulty on corresponding knowledge concepts. Therefore, the diagnostic function can be formulated as:

$$\mathbf{p}_{se}^{\tilde{\mathcal{G}}} = \mathbf{Q}_e \odot (\mathbf{a}_s - \mathbf{d}_e) \times h_e^{disc}, \ \hat{r}_{se}^{\tilde{\mathcal{G}}} = \sigma(MLPs(\mathbf{p}_{se}^{\tilde{\mathcal{G}}})),$$ (5)

where $\mathbf{Q}_e$ denotes the relations between concepts and exercise e, and $\odot$ denotes the element-wise product. $\mathbf{p}_{se}^{\tilde{\mathcal{G}}}$ denotes the hidden representation between student s and exercise e, and $\hat{r}_{se}^{\tilde{\mathcal{G}}}$ denotes the predicted student s's response log on exercise e based on graph $\tilde{\mathcal{G}}$. Finally, we use the classic BCE loss, formulated as:

$$\min_\theta \mathcal{L}_{BCE} = -\sum_{(s,e,r_{se}) \in R} (r_{se} \log(\hat{r}_{se}^{\tilde{\mathcal{G}}}) + (1 - r_{se}) \log(1 - \hat{r}_{se}^{\tilde{\mathcal{G}}})).$$ (6)

where θ denotes all trainable parameters of graph-based CD models, i.e., trainable parameters in Section 3.3. The remaining problem relies on how to obtain the reliable graph $\tilde{\mathcal{G}}$.

3.4 Informative Edge Differentiation Layer

In this part, we aim to build a reliable student-exercise graph $\tilde{\mathcal{G}}$ based on the original graph $\mathcal{G}$. To improve diagnostic performance, the reliable graph $\tilde{\mathcal{G}}$ should minimize the effects of uncertain edges. To this end, we propose an Informative Edge Differentiation (IE-Diff) layer to obtain this reliable graph. Our intuition is from a powerful information bottleneck principle [34, 45], which focuses on extracting relevant information from data while simultaneously discarding irrelevant details. Formally, given input data X, hidden representations H, and downstream task labels Y (chain form: $< X \rightarrow H \rightarrow Y >$), the information bottleneck principle suggests that optimal representations should keep minimal yet sufficient information for downstream tasks [29, 52], formulated as:

$$H^* = argmax_H I(Y; H) - \beta I(X; H), \tag{7}$$

where $I(Y; H)$ and $I(X; H)$ denote the mutual information between corresponding variables, and β denotes the balancing parameter between two parts. As shown in Figure 3, we adopt the information bottleneck principle in reducing effects of uncertain edges in the student-exercise bipartite graph, formulated as:

$$\max_{\phi} I(R; \mathbf{U}, \mathbf{V}, \tilde{\mathcal{G}}) - \beta I(\tilde{\mathcal{G}}, \mathcal{G}), \tag{8}$$

where R denotes response logs, $\mathbf{U}, \mathbf{V}$ denote latent embeddings of students and exercises. $\mathcal{G}$ and $\tilde{\mathcal{G}}$ denote the original bipartite graph and the reliable graph, respectively. ϕ denotes parameters of the reliable graph $\tilde{\mathcal{G}}$. Eq.(8) requires maximizing $I(R; \mathbf{A}, \mathbf{D}, \tilde{\mathcal{G}})$ and minimizing $I(\tilde{\mathcal{G}}, \mathcal{G})$. One obvious challenge in realization Eq.(8) lies in how to parameterize $\tilde{\mathcal{G}}$.

3.4.1 Parameterization of the reliable graph $\tilde{\mathcal{G}}$. In Eq.(8), $\tilde{\mathcal{G}}$ should be learnable. Typically, a student-exercise graph can be expressed as a binary adjacency matrix, while this matrix itself can not be optimized. Therefore, a key challenge lies in the parameterization of $\tilde{\mathcal{G}}$. In this paper, we formulate the edge parameterization as a graph edge dropout problem [52]. The detailed form is $\tilde{g}_{se} = g_{se} \odot \rho_{se}$, where g_{se} and $\tilde{g}_{se}$ denote the edge between student s and exercise e in the original graph $\mathcal{G}$ and reliable graph $\tilde{\mathcal{G}}$, respectively. $\odot$ denotes the element-wise product. ρ_{se} describes the certainty of edge g_{se}. The more uncertain the edge is, the higher the probability of being dropped. We model edge certainty for different subgraphs. Depending on whether edge g_{se} corresponds to a correct response log or an incorrect response log, we choose ρ_{se}^1 and ρ_{se}^0 to model its certainty, respectively. ρ_{se}^1 and ρ_{se}^0 together form the overall ρ_{se}. Then, we model them as $\rho_{se}^1 \sim Bern(w_{se}^1)$ and $\rho_{se}^0 \sim Bern(w_{se}^0)$. $Bern(w_{se}^1)$ and $Bern(w_{se}^0)$ represent Bernoulli distributions with parameter w_{se}^1 and w_{se}^0, respectively. Then, we propose to parameterize w_{se}^1 and w_{se}^0 with considering different semantics as:

$$w_{se}^0 = \mathbf{W}^0([\mathbf{u}_s; \mathbf{v}_e]), \ w_{se}^1 = \mathbf{W}^1([\mathbf{u}_s; \mathbf{v}_e]), \tag{9}$$

where $\mathbf{W}^0, \mathbf{W}^1$ denotes corresponding trainable weight matrices. $[;]$ denotes concatenation of two embeddings. After obtaining w_{se}^0 and w_{se}^1, we focus on making $Bern(w_{se}^0)$ and $Bern(w_{se}^1)$ differentiable. We adopt the popular concrete relaxation method [18]:

$$Bern(w_{se}^0) = \sigma(\log(\delta/(1-\delta) + w_{se}^0)/t),$$
$$Bern(w_{se}^1) = \sigma(\log(\delta/(1-\delta) + w_{se}^1)/t), \tag{10}$$

where t is the temperature parameter and we set $t = 0.2$ in our experiments. δ denotes random variable sampled from $U(0, 1)$.

3.4.2 Maximization of $I(R; \mathbf{U}, \mathbf{V}, \tilde{\mathcal{G}})$. In this part, we focus on how to handle the first term in Eq.(8). $I(R; \mathbf{U}, \mathbf{V}, \tilde{\mathcal{G}})$ requires that the reliable graph $\tilde{\mathcal{G}}$ should satisfy the response log prediction. Here, we derive the lower bound of $I(R; \mathbf{U}, \mathbf{V}, \tilde{\mathcal{G}})$ for its maximization.

$$I(R; \mathbf{U}, \mathbf{V}, \tilde{\mathcal{G}}) = H(R) - H(R|\mathbf{U}, \mathbf{V}, \tilde{\mathcal{G}})$$

$$\overset{(a)}{\geq} \sum_{s=1}^{M} \sum_{e=1}^{N} \sum_{r=0}^{1} p(r_{se}, \mathbf{u}_s, \mathbf{v}_e, \mathcal{G}) \log p(r_{se}|\mathbf{u}_s, \mathbf{v}_e, \tilde{\mathcal{G}})$$

$$\overset{(b)}{\geq} \sum_{(s,e,r_{se}) \in R} \log p(r_{se}|\mathbf{u}_s, \mathbf{v}_e, \tilde{\mathcal{G}})$$

$$\overset{(c)}{=} \sum_{(s,e,r_{se}) \in R} \log p(r_{se}|\hat{r}_{se}) \tag{11}$$

$$\overset{(d)}{=} \sum_{(s,e,r_{se}) \in R} \log[\hat{r}_{se}^{r_{se}}(1 - \hat{r}_{se})^{(1-r_{se})}]$$

$$\overset{(e)}{=} \sum_{(s,e,r_{se}) \in R} (r_{se} \log(\hat{r}_{se}) + (1 - r_{se}) \log(1 - \hat{r}_{se})).$$

Here, $\hat{r}_{se}$ is short for $\hat{r}_{se}^{\tilde{\mathcal{G}}}$. The reason for each derivation step is as follows: (a) is the non-negative property of entropy; (b) is that $p(r_{se}, \mathbf{u}_s, \mathbf{v}_e, \mathcal{G}) \leq 1$ and $\log p(r_{se}|\mathbf{u}_s, \mathbf{v}_e, \tilde{\mathcal{G}}) < 0$; (c) is that we can obtain $\hat{r}_{se}$ given students and exercises' parameters (this process is shown in Figure 2); (d) represents the form in the context of maximum likelihood probability; (e) is the property of *log* function. The lower bound of $I(R; \mathbf{A}, \mathbf{D}, \tilde{\mathcal{G}})$ is the negative of a Binary Cross-Entropy (BCE) function (a widely-adopted loss in CD models [39, 42]). Therefore, maximizing the mutual information is equivalent to minimizing the BCE loss in Eq.(6).

3.4.3 Minimization of $I(\tilde{\mathcal{G}}, \mathcal{G})$. Next, we focus on how to minimize the second term in Eq.(8). Estimating the upper bound of mutual information is difficult. Some methods employ variational techniques for this estimation [1–3], however, they heavily depend on prior assumptions. In this paper, we adopt the Hilbert-Schmidt Independence Criterion (**HSIC** [16, 25]) for mutual information minimization. $HSIC(X, Y)$ measures dependence between two variables X and Y. Generally, if $HSIC(X, Y)$ is closer to 0, these variables are more independent from each other. In detail, $HSIC(X, Y)$ is based on covariance operators in the Reproducing Kernel Hilbert Space [5], formulated as:

$$HSIC(X, Y) = \mathbb{E}_{X,X',Y,Y'}[\mathcal{K}_X(X, X')\mathcal{K}_Y(Y, Y')]$$
$$+ \mathbb{E}_{X,X'}[\mathcal{K}_X(X, X')]\mathbb{E}_{Y,Y'}[\mathcal{K}_Y(Y, Y')] \tag{12}$$
$$- 2\mathbb{E}_{XY}[\mathbb{E}_{X'}[\mathcal{K}_X(X, X')]\mathbb{E}_{Y'}[\mathcal{K}_Y(Y, Y')]],$$

where X' and Y' are two independent copies of X and Y. $\mathcal{K}_X$ and $\mathcal{K}_Y$ denote the kernel function for two variables. Given sufficient instances $\{x_i, y_i\}_{i=1}^n$ in the batched training samples, estimation of $HSIC(X, Y)$ can be realized as $HSIC(X, Y) = (n-1)^{-2} Tr(\mathcal{K}_X J \mathcal{K}_Y J)$, where $Tr(\cdot)$ denotes the trace of a matrix. $J = \mathbf{I} - \mathbf{11}^T/n$ denotes the centering matrix, where $\mathbf{I}$ denotes the Identity matrix, n denotes the number of batched instances [43]. The inputs of $\mathcal{K}_X$ and $\mathcal{K}_Y$ are two instances from batched samples, i.e., $\mathcal{K}_X(x_i, x_j)$ and $\mathcal{K}_Y(y_i, y_j)$. In this paper, we adopt a widely-adopted Radial Basis Function (RBF) for $\mathcal{K}_X$ [35]:

$$\mathcal{K}_X(x_i, x_j) = exp(-\frac{||x_i - x_j||^2}{2\alpha}), \tag{13}$$

where α denotes a constant that controls the sharpness of $\mathcal{K}_X$. Following [52], we set α to 0.2 in experiments. We also adopt the same RBF kernel function for $\mathcal{K}(y_i, y_j)$. In this paper, we aim to minimize mutual information between two graphs. Here, we utilize

the minimization of $HSIC(\tilde{\mathcal{G}}, \mathcal{G})$ to replace direct minimization of $I(\tilde{\mathcal{G}}, \mathcal{G})$. As graphs are non-Euclidean, we adopt Monte Carlo sampling on node representations for estimation [52]: $HSIC(\tilde{\mathcal{G}}, \mathcal{G})) = HSIC(\mathbf{U}^{\tilde{\mathcal{G}}}, \mathbf{U}^{\mathcal{G}}) + HSIC(\mathbf{V}^{\tilde{\mathcal{G}}}, \mathbf{V}^{\mathcal{G}})$, where $\mathbf{U}^{\tilde{\mathcal{G}}}$ and $\mathbf{V}^{\tilde{\mathcal{G}}}$ denote final embeddings based on the reliable graph $\tilde{\mathcal{G}}$. $\mathbf{U}^{\mathcal{G}}$ and $\mathbf{V}^{\mathcal{G}}$ denote embeddings based on the original graph $\mathcal{G}$. In our experiments, we take students and exercises in each batch into consideration. In summary, we realize minimizing mutual information via the following HSIC-based loss function:

$$L_{HSIC} = HSIC(\mathbf{U}^{\tilde{\mathcal{G}}}, \mathbf{U}^{\mathcal{G}}) + HSIC(\mathbf{V}^{\tilde{\mathcal{G}}}, \mathbf{V}^{\mathcal{G}}). \tag{14}$$

We combine Eq.(6) and Eq.(14) for the final loss, which can be formulated as:

$$\min_{\phi} \mathcal{L}_{all} = \mathcal{L}_{BCE} + \beta \mathcal{L}_{HSIC}, \tag{15}$$

where $\mathcal{L}_{all}$ is used to update parameters of IE-Diff (ϕ).

3.5 Alternating Training Strategy

A straightforward idea is to adopt the paradigm of multi-task learning and use Eq.(15) to optimize overall parameters of ISG-CD together (θ and ϕ) [52]. Considering the interdependence between the S-GNN based CD model and the IE-Diff layer, we adopt alternating training [58, 59]. Specifically, we choose to pre-train θ based on the original graph for a few epochs as a start. In practice, we allow for 5 additional pre-training epochs on all three datasets. Subsequently, we alternately perform two steps: fix θ to optimize ϕ (Eq.(15)), then fix ϕ to optimize θ (Eq.(6)) until convergence. To better highlight the advantages, we also compare different strategies in Section 4.6.

3.6 Model Discussions

3.6.1 Additional Time Complexity. Compared to the backbone model KaNCD, ISG-CD first involves with GNN. Specifically, we utilize sparse tensors to represent adjacency matrices. The Update of student/exercise latent embeddings relies on a sparse tensor multiplication, and the time complexity is $O(nnz * T)$, where T denotes the number of concepts and nnz represents the number of non-zero elements in the sparse matrix. Second, we introduce an informative edge differentiation layer. As shown in Eq.(10), this layer maps local embeddings to certainty for all edges. Note that, this process does not involve with GNNs, therefore, the time complexity caused by IE-Diff is much lower than S-GNN. Overall, the additional time complexity of ISG-CD is low.

3.6.2 Additional Space Complexity. ISG-CD has two parts of additional parameters compared to KaNCD. One part lies in adjacency matrices corresponding to subgraphs, whose space complexity of sparse matrices is $O(nnz)$. The other part is informative edge differentiation, which applies additional MLPs to calculate edge certainty. The additional MLPs are shared across all edges, therefore, space complexity of this part can be ignored. Overall, the additional space complexity of ISG-CD is low.

3.6.3 Training Procedures of ISG-CD. For readability, we present the training procedures of ISG-CD in Algorithm 1. The process is divided into updating the parameters of the graph structure and the parameters of the CD model sequentially within each epoch.

Algorithm 1 Detailed training procedures of ISG-CD.

Require: Triplets of students, exercises and response log R, student-exercise bipartite graph $\mathcal{G}$.

Ensure:

1: Initialize parameters of graph-based CD (θ) and parameters of edge differentiation layer (ϕ).
2: Pre-train θ based on the original graph $\mathcal{G}$ for several epochs.
3: **repeat**
4:     Fix θ and unfix ϕ;
5:     **for** each sample (s, e, r_{se}) in the training data **do**
6:         Obtain the reliable graph $\tilde{\mathcal{G}}$ (Eq.(10));
7:         Update $\mathbf{U}^{\tilde{\mathcal{G}}}, \mathbf{V}^{\tilde{\mathcal{G}}}$ according to $\tilde{\mathcal{G}}$ (Eq.(3));
8:         Update $\mathbf{U}^{\mathcal{G}}, \mathbf{V}^{\mathcal{G}}$ according to $\mathcal{G}$ (Eq.(3));
9:         Obtain student abilities and exercise difficulties based on matrix factorization (Eq.(4));
10:        Obtain response log prediction $\hat{r}_{se}^{\tilde{\mathcal{G}}}$ (Eq.(5));
11:        Calculate HSIC loss (Eq.14);
12:        Calculate BCE loss (Eq.(6));
13:        Calculate overall loss (Eq.(15));
14:     **end for**
15:     Minimize Eq.(15) to update ϕ,
16:     Fix ϕ and unfix θ;
17:     **for** each sample (s, e, r_{se}) in the training data **do**
18:         Obtain the reliable graph $\tilde{\mathcal{G}}$ (Eq.(10));
19:         Update $\mathbf{U}^{\tilde{\mathcal{G}}}, \mathbf{V}^{\tilde{\mathcal{G}}}$ according to $\tilde{\mathcal{G}}$ (Eq.(3));
20:         Obtain student abilities and exercise difficulties based on matrix factorization (Eq.(4));
21:         Obtain response log prediction $\hat{r}_{se}^{\tilde{\mathcal{G}}}$ (Eq.(5));
22:         Calculate BCE loss (Eq.(6));
23:     **end for**
24:     Minimize BCE loss to update θ,
25: **until** Convergence.

---

4 Experiments

In this section, we try to answer these Research Questions (RQ):

RQ1: Does ISG-CD have consistently superior performance on three datasets? (Section 4.2)
RQ2: Does ISG-CD really detect uncertain edges? (Section 4.3)
RQ3: Are all components in ISG-CD important? (Section 4.4)
RQ4: What are the impacts of different hyperparameters on ISG-CD? Can ISG-CD consistently achieve better diagnostic performance than the second-best CD model? (Section 4.5)
RQ5: What are the impacts of different training strategies on ISG-CD? (Section 4.6)

4.1 Experimental Settings

4.1.1 Datasets. In the experimental part, we choose three real-world datasets, i.e., ASSIST dataset (ASSISTments 2009-2010 "skill builder")[1], Junyi dataset (Junyi Academy Math Practicing Log)[2], and MOOC-Radar dataset[3]. ASSIST is a publicly available dataset

---

[1] https://sites.google.com/site/assistmentsdata/feng2009
[2] https://pslcdatashop.web.cmu.edu/DatasetInfo?datasetId=1198
[3] https://github.com/THU-KEG/MOOC-Radar

Table 1: The detailed statistics of three datasets.

Dataset	ASSIST	Junyi	MOOC-Radar
#Students	2,493	10,000	14,224
#Exercises	17,746	734	2,513
#Knowledge concepts	123	734	580
#Response logs	267,415	408,057	898,933
#Response logs per student	107.266	40.8	63.198
#Concepts per exercise	1.192	1	1
#Sparsity in response logs	99.396%	94.441%	97.485%

collected by the ASSISTments online tutoring systems [12]. It contains a wealth of student-exercise response logs and expert-labeled exercise-concept relations Q. Junyi dataset is an open dataset collected by an e-learning website, Junyi Academy. The unique characteristic of the Junyi dataset lies in one-to-one correspondence between knowledge concepts and exercises. We adopt the same pre-processing method as [19]. MOOC-Radar is a recently collected dataset from students' learning records in MOOCs [53], which contains abundant response logs. We record the detailed statistics in Table 1. Finally, we conduct five-fold cross-validation for all models. Specifically, we randomly split all student-exercise response logs into 5 parts. Each part will be treated as the testing set in turn, and then we split the remaining 4 parts into the training set and the validation set with the ratio of 7:1.

4.1.2 Metrics. Similar to previous studies [14, 19, 38, 42], we evaluate the prediction performance of response logs. Two commonly used accuracy metrics have been adopted for performance evaluation: Accuracy (ACC), and Area Under the Curve (AUC). Additionally, we employ a widely-used metric called Degree of Agreement (DOA) [39, 47]. DOA measures the consistency between the predicted proficiency levels and the observed patterns in response logs. Note that, we evaluate DOA on the testing set.

4.1.3 Baselines. We select the following cutting-edge and high-performing CD models as our baselines:

- **DINA** [10]. It adopts binary variables to represent students and exercises. Guess and slip parameters are also introduced.
- **NCDM** [38]. It applies neural networks to CD, and uses high-dimensional representations to represent students' abilities.
- **KaNCD** [39]. Compared to NCDM, it adopts matrix factorization techniques for student abilities and exercise difficulties.
- **KSCD** [22]. Compared to NCDM, it adopts matrix factorization techniques for student abilities and exercise difficulties, and designs a novel fusion function.
- **RCD** [14]. It incorporates student-exercise bipartite graph and concept-concept relationships into CD models.
- **SCD** [42]. Compared to RCD, it gives up diagnosing proficiency levels, and introduces a self-supervised learning based loss.
- **HAN-CD** [44]. To adapt HAN to CD, we replace initial node features and projection with free embeddings, and consider correct/incorrect student-exercise meta paths.

For the sake of feasibility of KSCD with 24G cuda memory, we change its fusion function by directly fusing representations at the second dimension on Junyi and MOOC-Radar datasets. For RCD, we replace the GNN propagation process in the original codes

with sparse tensor matrix multiplication, which improves efficiency while keeping a similar diagnostic performance. Besides, we present a variation of edge differentiation layer in Appendix A.1, and the corresponding results are recorded in Appendix B.

4.1.4 Hyperparameter Settings. We randomly sample 8,192 logs per batch, and search the learning rate from {0.0001, 0.0005, 0.001, 0.005, 0.01} for all models. We adopt the Adam optimizer. We set the embedding dimension Z to 128 for all models, and adopt the Xavier initialization for all trainable parameters. As for the number of GNN layers L, similar to previous studies [27, 42], we search from {1, 2, 3, 4} and select the best result. The hidden dimensions of MLPs in Eq.(5) are 512 and 256 for all models. Following [52], the temperature t is set to 0.2, α in Eq.(13) is set to 0.2 on all datasets. To ensure fair comparisons, for both ORCDF and our proposed ISG-CD, we choose KaNCD as the backbone model [39], and set the balancing hyperparameter β to 0.5. Further, we adopt the same GNN structure for these two models. We have released codes of ISG-CD for implementation[4].

Table 2: Overall performance on the ASSIST dataset. We use bold font to emphasize the best results and underline to indicate the second-best results.

Model	ACC ↑	AUC ↑	DOA ↑
DINA	0.6253 ± 0.0245	0.6794 ± 0.0201	0.5579 ± 0.0316
NCDM	0.7072 ± 0.0222	0.7244 ± 0.0212	0.5543 ± 0.0293
KSCD	0.7209 ± 0.0241	0.7503 ± 0.0252	0.5092 ± 0.0062
KaNCD	0.7182 ± 0.0250	0.7404 ± 0.0251	0.6057 ± 0.0235
RCD	0.7153 ± 0.0123	0.7382 ± 0.0226	0.6221 ± 0.0230
SCD	0.7212 ± 0.0566	<u>0.7552 ± 0.0576</u>	-
HAN-CD	<u>0.7257 ± 0.0229</u>	0.7524 ± 0.0240	<u>0.6348 ± 0.0185</u>
ISG-CD	**0.7322 ± 0.0247**	**0.7604 ± 0.0296**	**0.6582 ± 0.0251**

4.2 Overall Performance (RQ1)

We report overall performance in Table 2, Table 3 and Table 4. There are several observations from these three tables.

- **First of all, ISG-CD has the best performance on three datasets**. On three datasets, ISG-CD achieves a stable improvement in accuracy performance compared to second-best results. Furthermore, ISG-CD has an improvement of nearly 1.5% on the DOA metric on MOOC-Radar dataset. This clearly demonstrates the stable and strong modeling ability of ISG-CD.
- **Second, compared to graph-based CD models which do not distinguish edge heterogeneity (RCD, SCD), our proposed ISG-CD achieves stable improvements on all three datasets.** Specifically, compared to RCD and SCD, ISG-CD has over 1% accuracy improvements on ASSIST and Junyi datasets. This prove that distinguishing edge semantics in CD will improve diagnosis performance of graph-based CD models.
- **Third, we find that ISG-CD consistently outperforms both HAN-CD and ORCDF on three models**. ompared to HAN-CD and ORCDF, ISG-CD has nearly 0.7% AUC improvements

---

[4]https://github.com/ShaoPengyang/ISG-CD

Table 3: Overall performance on the Junyi dataset. We use bold font to emphasize the best results and underline to indicate the second-best results.

Model	ACC ↑	AUC ↑	DOA ↑
DINA	0.5105 ± 0.0230	0.6429 ± 0.0177	0.5002 ± 0.0056
NCDM	0.7482 ± 0.0013	0.7816 ± 0.0013	0.5476 ± 0.0162
KSCD	0.7561 ± 0.0029	0.7909 ± 0.0053	0.5001 ± 0.0039
KaNCD	0.7536 ± 0.0020	0.7867 ± 0.0017	0.5529 ± 0.0212
RCD	0.7512 ± 0.0044	0.7834 ± 0.0025	0.4996 ± 0.0050
SCD	0.7576 ± 0.0056	0.7902 ± 0.0040	-
HAN-CD	<u>0.7626 ± 0.0039</u>	<u>0.7957 ± 0.0080</u>	<u>0.6469 ± 0.0132</u>
ISG-CD	**0.7672 ± 0.0041**	**0.8058 ± 0.0039**	**0.6728 ± 0.0208**

Table 4: Overall performance on the MOOC-Radar dataset. We use bold font to emphasize the best results and underline to indicate the second-best results.

Model	ACC ↑	AUC ↑	DOA ↑
DINA	0.7792 ± 0.0047	0.8142 ± 0.0035	0.5436 ± 0.0141
NCDM	0.8537 ± 0.0063	0.8663 ± 0.0056	0.6055 ± 0.0247
KSCD	0.8587 ± 0.0060	0.8800 ± 0.0048	0.5027 ± 0.0139
KaNCD	0.8576 ± 0.0093	0.8782 ± 0.0049	0.6980 ± 0.0161
RCD	0.8599 ± 0.0062	0.8792 ± 0.0065	0.6678 ± 0.0121
SCD	<u>0.8617 ± 0.0068</u>	<u>0.8810 ± 0.0051</u>	-
HAN-CD	0.8600 ± 0.0037	0.8799 ± 0.0096	<u>0.7231 ± 0.0231</u>
ISG-CD	**0.8676 ± 0.0045**	**0.8885 ± 0.0068**	**0.7432 ± 0.0132**

on all three datasets. Note that, HAN-CD, ORCDF and our proposed ISG-CD actually share a similar GNN structure. These results indicate the effectiveness of constructing a reliable student-exercise graph structure in enhancing graph-based CD models.

- **Last but not least, we find that graph based methods (HAN-CD, ORCDF, and ISG-CD) achieve the most accuracy improvements on the Junyi dataset compared to the backbone KaNCD model**. The reason lies in the unique characteristic of Junyi dataset, i.e., one-to-one correspondence between knowledge concepts and exercises. That is to say, concepts that we encountered in the testing set will be totally different from those in the training set. This indicates that utilizing graph structure can better handle some extreme data situations.

4.3 Uncertain Edge Detection (RQ2)

In this part, we focus on whether the IE-Diff layer can accurately distinguish uncertain edges. Note that, these uncertain edges arise due to students' carelessness or guessing. We find that it is not easy to generate random noise on non-interacted student-exercise pairs. These non-interacted pairs consist of potential correct and incorrect response logs. Hence, we randomly choose certain ratios of existing student-exercise response logs (corresponding to edges in graph) and change their labels. IE-Diff assigns $Bern(w_{se}^0)$ or $Bern(w_{se}^1)$ to the edge between student s and exercise e. A smaller value of $Bern(w_{se}^0)$ or $Bern(w_{se}^1)$ denotes more uncertainty in this edge. If the value is smaller than 0.5, IE-Diff regards this edge as detected uncertain edges. We conduct experiments about whether these uncertain edges can be detected on ASSIST and Junyi datasets.
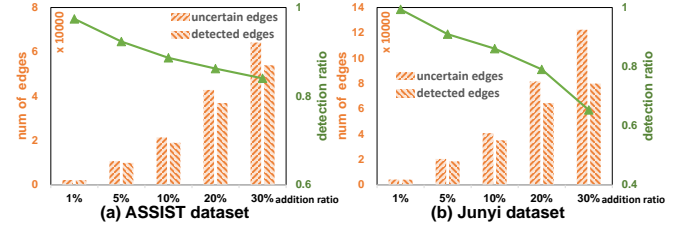

Figure 4: Edge detection with varying the number of uncertain edges on the ASSIST and Junyi datasets.
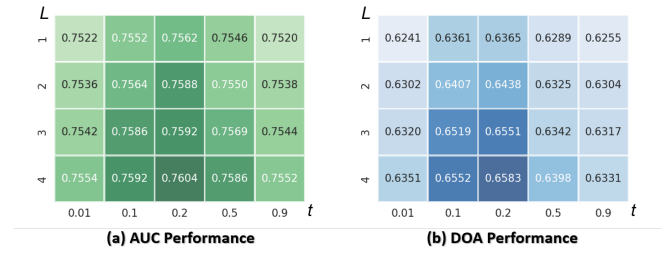

Figure 5: Diagnostic performance with varying number of GNN layers L and temperature t on the ASSIST dataset.

From Figure 4, we can infer two findings. First, the more modified edges, the lower the detection ratio. One possible reason is that the original data distribution will change if most edges are uncertain. Second, we find that even with a considerable number of uncertain edges, ISG-CD still maintains relatively reliable detection performance. Specifically, on the ASSIST dataset, even with 30% of the edges being uncertain, we still maintain a detection probability of near 80%. These findings can reflect effectiveness of IE-Diff.

4.4 Ablation Studies (RQ3)

We have conducted ablation experiments about the Semantic GNN (S-GNN) layer and Informative Edge Differentiation layer (IE-Diff) of ISG-CD, and results are presented in Table 5. The first row does not take any additional layers into consideration, therefore, it corresponds to the backbone KaNCD model [39]. Please note that, S-GNN layer serves as the precondition for IE-Diff layer, therefore, we do not include adopting IE-Diff without S-GNN in Table 5.

There are several observations from Table 5. First, Semantic-aware GNN Layer demonstrates remarkable adaptability to CD by combining state-of-the-art model paradigms [38, 39]. Compared to some models that sacrifice diagnostic capabilities for accuracy improvements [42], it not only achieves a remarkable accuracy improvement but also retains its diagnostic capabilities intact. Second, Informative Edge Differentiation Layer makes precise adjustments to the graph structure, resulting in consistent accuracy improvements on both datasets. This layer builds on the aforementioned foundations, showcasing its ability to further enhance diagnostic performance.

Table 5: Ablation studies on ASSIST and Junyi datasets. The last row corresponds to our proposed ISG-CD.

S-GNN	IE-Diff	ASSIST dataset			Junyi dataset		
Sec 3.3	Sec 3.4	ACC ↑	AUC ↑	DOA ↑	ACC ↑	AUC ↑	DOA ↑
✗	✗	0.7182 ± 0.0250	0.7404 ± 0.0251	0.6057 ± 0.0231	0.7536 ± 0.0020	0.7867 ± 0.0017	0.5529 ± 0.0212
✓	✗	0.7248 ± 0.0165	0.7518 ± 0.0188	0.6329 ± 0.0101	0.7623 ± 0.0082	0.7964 ± 0.0065	0.6432 ± 0.0123
✓	✓	$\mathbf{0.7322 \pm 0.0247}$	$\mathbf{0.7604 \pm 0.0296}$	$\mathbf{0.6582 \pm 0.0251}$	$\mathbf{0.7672 \pm 0.0041}$	$\mathbf{0.8058 \pm 0.0041}$	$\mathbf{0.6728 \pm 0.0208}$
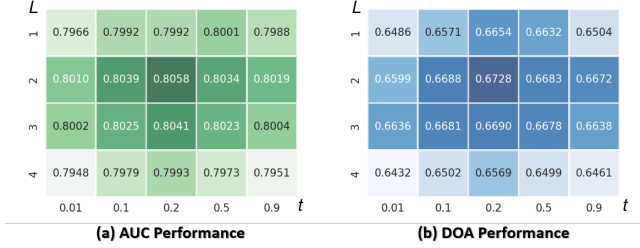
Figure 6: Diagnostic performance with varying number of GNN layers L and temperature t on the Junyi dataset.

4.5 Hyperparameters Analyses (RQ4)

Number of GNN Layers L and Temperature t. The number of GNN layers L is directly related to performance of ISG-CD. Also, temperature t in Eq.(10) is an important hyperparameter that controls the effectiveness of $\mathcal{L}_{HSIC}$. In this part, we focus on studying the impacts of these two factors. The results are in Figure 5-6.

We have several observations from these results. First, we find that ISG-CD achieves a stable improvement compared to the backbone model, KaNCD. Specifically, AUC of KaNCD on ASSIST dataset is 0.7404. However, ACC of ISG-CD in Figure 5 is over 0.7550 with ranging the number of GNN layers L from 1 to 4. This indicates the stable effectiveness of our proposed semantic-aware GNN layers. Second, we find that the optimal results on the ASSIST dataset occur when $L = 4$, while the optimal results on the Junyi dataset occur when $L = 2$. The phenomenon on Junyi dataset indicates that there may be over-smoothing caused by multiple GNN layers. These findings suggest that the number of GNN layers should be carefully chosen based on specific datasets. Third, when the temperature t is in the range of 0.1 to 0.5, ISG-CD achieves relatively stable performance on both two datasets. When the temperature t approaches extreme values, ISG-CD experiences a noticeable decline in diagnosis performance. For example, on the ASSIST dataset, AUC of ISG-CD is close to 0.76 when $t = 0.1, 0.2, 0.5$ and $L = 4$, while AUC is close to 0.755 when $t = 0.01, 0.9$ and $L = 4$. These observation suggests that we should avoid extreme values when choosing temperature t.

4.6 Comparing Different Training Strategies (RQ 5)

We compare the following training strategies: the Alternating Learning (AL) strategy [58, 59] and the Multi-Task Learning (MTL) strategy [52]. For the process of the AL strategy, we first pre-train θ based on the original graph for 5 epochs. We also compare AL with ALwoP (AL without Pre-training). Subsequently, we alternately

Table 6: Comparing Different training strategies on the ASSIST and Junyi datasets.

Dataset	strategy	ACC ↑	AUC ↑	DOA ↑
ASSIST	MTL	0.7271 ± 0.0186	0.7532 ± 0.0160	0.6299 ± 0.0211
ASSIST	ALwoP	0.7301 ± 0.0284	0.7586 ± 0.0311	0.6402 ± 0.0268
ASSIST	AL	0.7322 ± 0.0247	0.7604 ± 0.0296	0.6582 ± 0.0251
Junyi	MTL	0.7621 ± 0.0066	0.8004 ± 0.0052	0.6524 ± 0.0137
Junyi	ALwoP	0.7646 ± 0.0052	0.8028 ± 0.0038	0.6633 ± 0.0162
Junyi	AL	0.7672 ± 0.0041	0.8058 ± 0.0039	0.6728 ± 0.0208

perform two steps: fix θ to optimize ϕ (Eq.(15)), then fix ϕ to optimize θ (Eq.(6)) until convergence. As for the process of the MTL strategy, we directly use Eq.(15) to optimize overall parameters of ISG-CD together (θ and ϕ) [52].

The corresponding results are in Table 6. Firstly, the performance of MTL is the worst. A possible reason is that the parameters of the graph-based CD models (θ) do not need to be constrained by HSIC loss, but MTL forces all parameters to be constrained by both BCE and HSIC. Secondly, we find that the pre-training stage can improve the performance of ISG-CD. We believe that the pre-training stage can enhance the stability of ISG-CD by providing better initial values of θ for the following alternating training.

5 Conclusion and Future Work

Educational cognitive diagnosis is a critical task in the data mining domain, which focuses on diagnosing students' proficiency levels on knowledge concepts. In this paper, we proposed a novel ISG-CD, which captures the edge heterogeneity and uncertainty. Firstly, to capture edge heterogeneity, we designed the process of how to incorporate a semantic-aware GNN into CD. Secondly, we proposed an Informative Edge Differentiation layer, which focused on keeping a minimal yet sufficient reliable graph for CD based on the information bottleneck principle. Specifically, we formulated the process as goals of mutual information minimization/maximization, and converted these goals to different loss functions. Finally, we adopted alternating training to optimize the graph-based CD model and edge differentiation layer. Extensive experimental results on three datasets demonstrated the superiority of ISG-CD. In the future, we plan to explore other solutions to alleviating negative effects of uncertain edges.

Acknowledgments

References

[1] Eslam Abdelaleem, Ilya Nemenman, and K Michael Martini. 2023. Deep Variational Multivariate Information Bottleneck–A Framework for Variational Losses. *arXiv preprint arXiv:2310.03311* (2023).

[2] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. 2016. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410* (2016).

[3] Seojin Bang, Pengtao Xie, Heewook Lee, Wei Wu, and Eric Xing. 2021. Explaining a black-box by using a deep variational information bottleneck approach. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 11396–11404.

[4] Rianne van den Berg, Thomas N Kipf, and Max Welling. 2017. Graph convolutional matrix completion. *arXiv preprint arXiv:1706.02263* (2017).

[5] Alain Berlinet and Christine Thomas-Agnan. 2011. *Reproducing kernel Hilbert spaces in probability and statistics.* Springer Science & Business Media.

[6] Miaomiao Cai, Lei Chen, Yifan Wang, Haoyue Bai, Peijie Sun, Le Wu, Min Zhang, and Meng Wang. 2024. Popularity-aware alignment and contrast for mitigating popularity bias. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining.* 187–198.

[7] Jiawei Chen, Junkang Wu, Jiancan Wu, Xuezhi Cao, Sheng Zhou, and Xiangnan He. 2023. Adap-τ: Adaptively Modulating Embedding Magnitude for Recommendation. In *Proceedings of the ACM Web Conference 2023.* 1085–1096.

[8] Lei Chen, Le Wu, Kun Zhang, Richang Hong, Defu Lian, Zhiqiang Zhang, Jun Zhou, and Meng Wang. 2023. Improving recommendation fairness via data augmentation. In *Proceedings of the ACM Web Conference 2023.* 1012–1020.

[9] Jiajun Cui, Hong Qian, Bo Jiang, and Wei Zhang. 2024. Leveraging Pedagogical Theories to Understand Student Learning Process with Graph-based Reasonable Knowledge Tracing. *arXiv preprint arXiv:2406.12896* (2024).

[10] Jimmy De La Torre. 2009. DINA model and parameter estimation: A didactic. *Journal of educational and behavioral statistics* 34, 1 (2009), 115–130.

[11] Susan E Embretson and Steven P Reise. 2013. *Item response theory.* Psychology Press.

[12] Mingyu Feng, Neil Heffernan, and Kenneth Koedinger. 2009. Addressing the assessment challenge with an online system that tutors as it assesses. *User modeling and user-adapted interaction* 19 (2009), 243–266.

[13] Chongming Gao, Shiqi Wang, Shijun Li, Jiawei Chen, Xiangnan He, Wenqiang Lei, Biao Li, Yuan Zhang, and Peng Jiang. 2023. CIRS: Bursting filter bubbles by counterfactual interactive recommender system. *ACM Transactions on Information Systems* 42, 1 (2023), 1–27.

[14] Weibo Gao, Qi Liu, Zhenya Huang, Yu Yin, Haoyang Bi, Mu-Chun Wang, Jianhui Ma, Shijin Wang, and Yu Su. 2021. RCD: Relation map driven cognitive diagnosis for intelligent education systems. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval.* 501–510.

[15] Weibo Gao, Hao Wang, Qi Liu, Fei Wang, Xin Lin, Linan Yue, Zheng Zhang, Rui Lv, and Shijin Wang. 2023. Leveraging Transferable Knowledge Concept Graph Embedding for Cold-Start Cognitive Diagnosis. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 983–992.

[16] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. 2005. Measuring statistical dependence with Hilbert-Schmidt norms. In *International conference on algorithmic learning theory.* Springer, 63–77.

[17] Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. 2020. Heterogeneous graph transformer. In *Proceedings of the web conference 2020.* 2704–2710.

[18] Eric Jang, Shixiang Gu, and Ben Poole. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144* (2016).

[19] Jiatong Li, Fei Wang, Qi Liu, Mengxiao Zhu, Wei Huang, Zhenya Huang, Enhong Chen, Yu Su, and Shijin Wang. 2022. HierCDF: A Bayesian Network-based Hierarchical Cognitive Diagnosis Framework. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining.* 904–913.

[20] Shuo Liu, Junhao Shen, Hong Qian, and Aimin Zhou. 2024. Inductive Cognitive Diagnosis for Fast Student Learning in Web-Based Intelligent Education Systems. In *Proceedings of the ACM on Web Conference 2024.* 4260–4271.

[21] Xinpu Liu, Yanxin Ma, Ke Xu, Jianwei Wan, and Yulan Guo. 2022. AGFA-Net: Adaptive global feature augmentation network for point cloud completion. *IEEE Geoscience and Remote Sensing Letters* 19 (2022), 1–5.

[22] Haiping Ma, Manwei Li, Le Wu, Haifeng Zhang, Yunbo Cao, Xingyi Zhang, and Xuemin Zhao. 2022. Knowledge-Sensed Cognitive Diagnosis for Intelligent Education Platforms. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management.* 1451–1460.

[23] Haiping Ma, Siyu Song, Chuan Qin, Xiaoshan Yu, Limiao Zhang, Xingyi Zhang, and Hengshu Zhu. 2024. DGCD: An Adaptive Denoising GNN for Group-level Cognitive Diagnosis. In *The 33rd International Joint Conference on Artificial Intelligence (IJCAI-24).*

[24] Haiping Ma, Changqian Wang, Hengshu Zhu, Shangshang Yang, Xiaoming Zhang, and Xingyi Zhang. 2024. Enhancing cognitive diagnosis using un-interacted exercises: A collaboration-aware mixed sampling approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 8877–8885.

[25] Wan-Duo Kurt Ma, JP Lewis, and W Bastiaan Kleijn. 2020. The HSIC bottleneck: Deep learning without back-propagation. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 5085–5092.

[26] Andriy Mnih and Russ R Salakhutdinov. 2007. Probabilistic matrix factorization. *Advances in neural information processing systems* 20 (2007).

[27] Hong Qian, Shuo Liu, Mingjia Li, Bingdong Li, Zhi Liu, and Aimin Zhou. 2024. ORCDF: An Oversmoothing-Resistant Cognitive Diagnosis Framework for Student Learning in Online Education Systems. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining.* 2455–2466.

[28] Mark D Reckase. 1997. The past and future of multidimensional item response theory. *Applied Psychological Measurement* 21, 1 (1997), 25–36.

[29] Andrew M Saxe, Yamini Bansal, Joel Dapello, Madhu Advani, Artemy Kolchinsky, Brendan D Tracey, and David D Cox. 2019. On the information bottleneck theory of deep learning. *Journal of Statistical Mechanics: Theory and Experiment* 2019, 12 (2019), 124020.

[30] Pengyang Shao, Kun Zhang, Chen Gao, Lei Chen, Miaomiao Cai, Le Wu, Yong Li, and Meng Wang. [n. d.]. Breaking Student-Concept Sparsity Barrier for Cognitive Diagnosis. *Frontiers of Computer Science* ([n. d.]). doi:10.1007/s11704-025-40591-2

[31] Junhao Shen, Hong Qian, Shuo Liu, Wei Zhang, Bo Jiang, and Aimin Zhou. 2024. Capturing Homogeneous Influence among Students: Hypergraph Cognitive Diagnosis for Intelligent Education Systems. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining.* 2628–2639.

[32] Lichao Sun, Yingtong Dou, Carl Yang, Kai Zhang, Ji Wang, S Yu Philip, Lifang He, and Bo Li. 2022. Adversarial attack and defense on graph data: A survey. *IEEE Transactions on Knowledge and Data Engineering* 35, 8 (2022), 7693–7711.

[33] Qingyun Sun, Jianxin Li, Hao Peng, Jia Wu, Xingcheng Fu, Cheng Ji, and S Yu Philip. 2022. Graph structure learning with variational information bottleneck. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 4165–4174.

[34] Naftali Tishby and Noga Zaslavsky. 2015. Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw).* IEEE, 1–5.

[35] Jean-Philippe Vert, Koji Tsuda, and Bernhard Schölkopf. 2004. A primer on kernel methods. (2004).

[36] Fei Wang, Weibo Gao, Qi Liu, Jiatong Li, Guanhao Zhao, Zheng Zhang, Zhenya Huang, Mengxiao Zhu, Shijin Wang, Wei Tong, et al. 2024. A survey of models for cognitive diagnosis: New developments and future directions. *arXiv preprint arXiv:2407.05458* (2024).

[37] Fei Wang, Zhenya Huang, Qi Liu, Enhong Chen, Yu Yin, Jianhui Ma, and Shijin Wang. 2023. Dynamic Cognitive Diagnosis: An Educational Priors-Enhanced Deep Knowledge Tracing Perspective. *IEEE Transactions on Learning Technologies* (2023).

[38] Fei Wang, Qi Liu, Enhong Chen, Zhenya Huang, Yuying Chen, Yu Yin, Zai Huang, and Shijin Wang. 2020. Neural cognitive diagnosis for intelligent education systems. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 6153–6161.

[39] Fei Wang, Qi Liu, Enhong Chen, Zhenya Huang, Yu Yin, Shijin Wang, and Yu Su. 2022. NeuralCD: a general framework for cognitive diagnosis. *IEEE Transactions on Knowledge and Data Engineering* (2022).

[40] Jihong Wang, Minnan Luo, Jundong Li, Ziqi Liu, Jun Zhou, and Qinghua Zheng. 2023. Toward enhanced robustness in unsupervised graph representation learning: A graph information bottleneck perspective. *IEEE Transactions on Knowledge and Data Engineering* (2023).

[41] Shanshan Wang, Zhen Zeng, Xun Yang, Ke Xu, and Xingyi Zhang. 2024. Boosting neural cognitive diagnosis with student's affective state modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 620–627.

[42] Shanshan Wang, Zhen Zeng, Xun Yang, and Xingyi Zhang. 2023. Self-supervised Graph Learning for Long-tailed Cognitive Diagnosis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 110–118.

[43] Tinghua Wang, Xiaolu Dai, and Yuze Liu. 2021. Learning with Hilbert–Schmidt independence criterion: A review and new perspectives. *Knowledge-based systems* 234 (2021), 107567.

[44] Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Yanfang Ye, Peng Cui, and Philip S Yu. 2019. Heterogeneous graph attention network. In *The world wide web conference.* 2022–2032.

[45] Zifeng Wang, Tong Jian, Aria Masoomi, Stratis Ioannidis, and Jennifer Dy. 2021. Revisiting hilbert-schmidt information bottleneck for adversarial robustness. *Advances in Neural Information Processing Systems* 34 (2021), 586–597.

[46] Chenwang Wu, Xiting Wang, Defu Lian, Xing Xie, and Enhong Chen. 2023. A causality inspired framework for model interpretation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining.* 2731–2741.

[47] Le Wu, Xiangzhi Chen, Fei Liu, Junsong Xie, Chenao Xia, Zhengtao Tan, Mi Tian, Jinglong Li, Kun Zhang, Defu Lian, Hong Richang, and Wang Meng. 2024. EduStudio: towards a unified library for student cognitive modeling. *Frontiers of Computer Science* 19, 8 (2024), 198342.

[48] Le Wu, Yonghui Yang, Kun Zhang, Richang Hong, Yanjie Fu, and Meng Wang. 2020. Joint item recommendation and attribute inference: An adaptive graph convolutional network approach. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval.* 679–688.

[49] Tailin Wu, Hongyu Ren, Pan Li, and Jure Leskovec. 2020. Graph information bottleneck. *Advances in Neural Information Processing Systems* 33 (2020), 20437–20448.

[50] Shangshang Yang, Mingyang Chen, Ziwen Wang, Xiaoshan Yu, Panpan Zhang, Haiping Ma, and Xingyi Zhang. 2024. DisenGCD: A Meta Multigraph-assisted Disentangled Graph Learning Framework for Cognitive Diagnosis. *arXiv preprint arXiv:2410.17564* (2024).

[51] Yonghui Yang, Le Wu, Richang Hong, Kun Zhang, and Meng Wang. 2021. Enhanced graph learning for collaborative filtering via mutual information maximization. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 71–80.

[52] Yonghui Yang, Le Wu, Zihan Wang, Zhuangzhuang He, Richang Hong, and Meng Wang. 2024. Graph Bottlenecked Social Recommendation. *arXiv preprint arXiv:2406.08214* (2024).

[53] Jifan Yu, Mengying Lu, Qingyang Zhong, Zijun Yao, Shangqing Tu, Zhengshan Liao, Xiaoya Li, Manli Li, Lei Hou, Hai-Tao Zheng, et al. 2023. Moocradar: A fine-grained and multi-aspect knowledge repository for improving cognitive student modeling in moocs. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2924–2934.

[54] Dacao Zhang, Kun Zhang, Le Wu, Mi Tian, Richang Hong, and Meng Wang. 2024. Path-Specific Causal Reasoning for Fairness-aware Cognitive Diagnosis. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 4143–4154.

[55] Yifei Zhang, Hao Zhu, Ziqiao Meng, Piotr Koniusz, and Irwin King. 2022. Graph-adaptive rectified linear unit for graph neural networks. In *Proceedings of the ACM Web Conference 2022*. 1331–1339.

[56] Zheng Zhang, Qi Liu, Hao Jiang, Fei Wang, Yan Zhuang, Le Wu, Weibo Gao, and Enhong Chen. 2023. Fairlisa: Fair user modeling with limited sensitive attributes information. In *Thirty-seventh Conference on Neural Information Processing Systems*.

[57] Zheng Zhang, Le Wu, Qi Liu, Jiayu Liu, Zhenya Huang, Yu Yin, Yan Zhuang, Weibo Gao, and Enhong Chen. 2024. Understanding and improving fairness in cognitive diagnosis. *Science China Information Sciences* 67, 5 (2024), 152106.

[58] Yaochen Zhu, Liang Wu, Qi Guo, Liangjie Hong, and Jundong Li. 2024. Collaborative large language model for recommender systems. In *Proceedings of the ACM on Web Conference 2024*. 3162–3172.

[59] Marcelo Victor Wust Zibetti, Florian Knoll, and Ravinder R Regatte. 2022. Alternating learning approach for variational networks and undersampling pattern in parallel MRI applications. *IEEE transactions on computational imaging* 8 (2022), 449–461.

A   More Details

A.1   Variation of Edge Differentiation: Ada-Diff

Edge differentiation focuses on how to distinguish fortunate guesses (careless mistakes) from mastery (no mastery). This is crucial for GNNs because these uncertain edges may propagate incorrect information during aggregation. We have introduced IE-Diff based on the information bottleneck principle. Here, we introduce an intuitive approach, which is based on *A*daptive Edge *Diff*erentiation (**Ada-Diff**), formulated as:

$$\tilde{\mathcal{G}}_1 = \{(s,e)|\hat{r}_{se}^{\tilde{\mathcal{G}}} > \Delta \wedge r_{se} = 1\}, \tilde{\mathcal{G}}_0 = \{(s,e)|\hat{r}_{se}^{\tilde{\mathcal{G}}} < 1 - \Delta \wedge r_{se} = 0\}, \quad (16)$$

where Δ denotes the threshold for adaptively distinguishing uncertain edges. For a specific concept, if a student answers corresponding exercises correctly multiple times and only answers incorrectly once, the incorrect answer will be considered as an uncertain edge.

A.2   Notation Table

For readability, we list important notations in Table 7.

B   More Experimental Results

B.1   Comparisons between IE-Diff and Ada-Diff

In this part, we focus on comparing the performance of Ada-Diff and IE-Diff. We present the results in Table 8.

We have two observations from Table 8. First, Ada-Diff has a slightly worse performance than IE-Diff, though Ada-Diff has relatively superior performance than some strong baselines. Second, Ada-Diff necessitates an additional inference step on all response

Table 7: Important Notations in ISG-CD.

Task Formulation	
S, E, K	student, exercise, concept entities
M, N, T	number of students, exercises, concepts
r_{se}	students's response log to exercise e
$R = (s, e, r_{se})$	triplets of students, exercises and response logs
$\mathbf{Q} = \{q_{ek}\}^{N \times T}$	relations between exercises and concepts
$\mathbf{A}, \mathbf{a}_s$	student abilities, student s's ability
$\mathbf{D}, \mathbf{d}_e$	exercise difficulties, exercise e's difficulty
a_{sk}	student s's ability on concept k
d_{ek}	exercise e's difficulty on concept k
$\mathbf{h}^{disc}, h_e^{disc}$	exercise discriminations, exercise e's discrimination
Semantic-aware GNN (S-GNN) based CD	
$\mathbf{U}, \mathbf{u}_s = \mathbf{u}_s^{(0)}$	student embeddings, student s's local embedding
$\mathbf{V}, \mathbf{v}_e = \mathbf{v}_e^{(0)}$	exercise embeddings, exercise e's local embedding
$\mathbf{O}, \mathbf{o}_k$	concept embeddings, concept k's embedding
Z	the dimension of latent embeddings
$\mathcal{G}$	the original student-exercise graph
$\tilde{\mathcal{G}}$	the reliable student-exercise graph
$\tilde{\mathcal{G}}_1, \tilde{\mathcal{G}}_0$	student-exercise subgraphs with different edge types
$\mu_{e\to s, \tilde{\mathcal{G}}_1}^{(l)}, \mu_{e\to s, \tilde{\mathcal{G}}_0}^{(l)}$	edge-specific messages from e to s at l-th layer
$\mu_{s\to e, \tilde{\mathcal{G}}_1}^{(l)}, \mu_{s\to e, \tilde{\mathcal{G}}_0}^{(l)}$	edge-specific messages from s to e at l-th layer
$\mathcal{N}_{s, \tilde{\mathcal{G}}_1}, \mathcal{N}_{s, \tilde{\mathcal{G}}_0}$	student s's neighbor sets in different subgraphs
$\mathcal{N}_{e, \tilde{\mathcal{G}}_1}, \mathcal{N}_{e, \tilde{\mathcal{G}}_0}$	exercise e's neighbor sets in different subgraphs
$\mathbf{u}_s^{\tilde{\mathcal{G}}}, \mathbf{v}_e^{\tilde{\mathcal{G}}}$	final embeddings of student s and exercise e
$<, >; \sigma$	inner dot and sigmoid activation
$\hat{r}_{se}^{\tilde{\mathcal{G}}}$	predicted response log based on graph $\tilde{\mathcal{G}}$
$\mathbf{p}_{se}^{\tilde{\mathcal{G}}}$	hidden representation based on graph $\tilde{\mathcal{G}}$
θ	all learnable parameters of S-GNN based CD
Informative Edge Differentiation (IE-Diff) layer	
$I(X; Y)$	mutual information between two variables
$g_{se}, \tilde{g}_{se}$	edge in the original/reliable graph
$(1 - \rho_{se})$	edge uncertainty (probability of edge dropping)
$\mathbf{W}_0, \mathbf{W}_1$	trainable weight matrices
w_{se}^0, w_{se}^1	parameters of Bernoulli distributions
$Bern(w_{se}^0), Bern(w_{se}^1)$	Bernoulli distributions for different edges
δ, t, ϵ	random variable, temperature, noise term
X', Y'	two independent copies of variables X, Y
$\mathcal{K}_X, \mathcal{K}_Y$	kernel functions of variable X, Y
J	the centering matrix
$Tr(\cdot)$	the trace of a matrix
$\mathbf{I}$	the identity matrix
β	balancing parameter between two tasks
α	sharpness of kernel functions
$\mathbf{U}^{\tilde{\mathcal{G}}}, \mathbf{V}^{\tilde{\mathcal{G}}}$	final embeddings based on the reliable graph
$\mathbf{U}^{\mathcal{G}}, \mathbf{V}^{\mathcal{G}}$	final embeddings based on the original graph
ϕ	all learnable parameters of IE-Diff

Table 8: Comparisons between IE-Diff and Ada-Diff on the ASSIST and Junyi datasets.

Data	Model	ACC ↑	AUC ↑	DOA ↑
ASSIST	**Ada-Diff**	0.7283 ± 0.0222	0.7555 ± 0.0244	0.6383 ± 0.0207
ASSIST	**IE-Diff**	0.7322 ± 0.0247	0.7604 ± 0.0296	0.6582 ± 0.0251
Junyi	**Ada-Diff**	0.7647 ± 0.0047	0.7998 ± 0.0061	0.6484 ± 0.0146
Junyi	**IE-Diff**	0.7672 ± 0.0041	0.8058 ± 0.0039	0.6728 ± 0.0208

logs at each epoch. IE-Diff only calculates the certainty of each edge based on student and exercise local latent embeddings. This process does not require an additional GNN process. Third, the performance of Ada-Diff heavily relies on the manual choice of threshold Δ. In summary, IE-Diff not only delivers superior performance metrics but also operates with a more efficient computational complexity.