DOI: https://doi.org/10.67203/jossme.2026.h4x5ddy3

Journal of Studies in Science and Mathematics Education

ISSN: 2630-662x

Published: 21.11.2026

Volume (issue): 6(2)

Page: 373 – 384

Hybrid Knowledge-Learner Joint Recommendation for Learning Partner Matching: Cold-Start Handling, Popularity Bias, and the Limits of Hard Contextual Filtering

Muhammad Mustapha Muhammad[1]*, Abba Almu[1], Habiba Muhammad Sani[1], Audu Ahmed[1], Abdullahi Usman Gulumbe[2]

[1]Usmanu Danfodiyo University Sokoto

[2]Federal University Birnin Kebbi

***Corresponding author:** elmustty@gmail.com

Abstract

Learning partner recommendation systems help online learners find compatible study partners, but graph-based approaches such as the Knowledge-Learner Joint Recommendation (KLJRec) system struggle with cold-start learners who have no interaction history and with popularity bias toward well-connected learners. This study evaluates a hybrid extension of a graph-based recommender that combines a GraphSAGE-style Graph Neural Network (GNN), a Profile Encoder that provides embeddings for learners without graph edges, contextual filtering on semester, major and preferred study time, and Maximum Marginal Relevance (MMR) re-ranking. Evaluation used a dataset of 206 learner records and 400 partnership edges privately collected from a distance learning institution, with ground truth defined as held-out test edges over five random 80/20 edge splits (mean ± standard deviation reported). Two flaws in the original evaluation were identified and corrected: a circular ground truth that guaranteed near-perfect precision, and an embedding objective without negative sampling that collapsed all embeddings. Under the corrected protocol, the original hard contextual filter reduced candidate lists to under two learners on average and lowered Recall@10 from about 0.04 to 0.003, a failure that was resolved by replacing the filter with a continuous contextual-compatibility score. With this change, the hybrid pipeline with MMR was statistically indistinguishable from a GNN-only baseline on Precision, Recall, F1 and NDCG (for example, Recall@10 of 0.040 versus 0.041) while lowering the Popularity Bias Index (2.98 versus 3.30). However, none of the learned models reliably outperformed random or popularity-only recommendation on this dataset, indicating that the available attributes and 320 training edges carry little learnable partnership signal. The study contributes a corrected evaluation protocol, evidence that exact-match contextual filtering is harmful, and an honest negative result that motivates validation on larger partnership datasets.

Keywords: learning partner recommendation, graph neural network, cold-start problem, popularity bias, maximum marginal relevance

1 Introduction

The rapid growth of online learning platforms and massive open online courses (MOOCs) has increasingly exposed learners to the challenge of finding suitable learning partners who share similar goals, academic interests, and learning preferences (Shou et al., 2023), a difficulty that became more pronounced during the COVID-19 pandemic. The physical and temporal separation between educators and learners in online settings often leads to feelings of isolation and powerlessness, prompting discontinuation of online courses. This has motivated the development of learning peer systems that build social networks from student activity data and suggest learning peers based on network link relationships. Nguyen (2020) argued that the unique qualities of individual students strongly influence the development of online learning communities, in which collaboration through group projects and conversation helps students accomplish shared objectives. Effective peer-matching has been shown to improve learning outcomes, motivation, and social engagement, yet traditional educational recommender systems mainly focus on learning content or resources rather than compatible learning partners.

Among systems addressing this gap, Shou et al. (2024) proposed a joint recommendation algorithm for knowledge concepts and learning partners — the Knowledge-Learner Joint Recommendation (KLJRec) model — based on an improved Multi-gate Mixture-of-Experts (MMoE) architecture. A heterogeneous information network (HIN) was constructed from a MOOC platform, with meta-paths selected to extract human–computer and student–student interaction information. While KLJRec outperformed several baseline and hybrid models in precision, recall, and diversity, two limitations persisted: (i)

new students without partner history could not be effectively recommended, because embeddings relied solely on existing social edges (the cold-start problem), and (ii) the system disproportionately favoured highly connected students, resulting in repeated recommendations for popular learners and reduced visibility for less-connected individuals (popularity bias).

This study examines whether a hybrid extension of a graph-based learning partner recommender can mitigate cold-start exclusion and popularity bias, and evaluates it under a non-circular protocol. The extension combines a Graph Neural Network (GNN) with a Profile Encoder that supplies embeddings from semester, major, and preferred study time for learners without graph edges, contextual compatibility scoring, and Maximum Marginal Relevance (MMR) re-ranking. The research questions were: RQ1, does contextual filtering and re-ranking improve recommendation accuracy over a GNN-only baseline when evaluated against held-out partnership edges; RQ2, does the hybrid design help cold-start learners; RQ3, does MMR re-ranking reduce popularity bias; and RQ4, do the learned models outperform trivial random and popularity-only recommenders. The study is limited to algorithm design, model training, and offline evaluation using a dataset of 206 learner records collected privately from a distance learning institution, and does not include live deployment on an operational e-learning platform. Graph-based partner recommendation is only useful for entire cohorts if it can serve learners who join without any collaboration record, and if it avoids concentrating recommendations on a small set of highly connected learners, which can discourage participation among less-connected or newly enrolled students. Hybrid designs that add profile-based representations, contextual constraints and diversity-aware re-ranking are a natural response to both problems, but their benefits are only meaningful if they are measured against ground truth that is independent of the recommender's own constraints. By testing such a design under a held-out-edge protocol with trivial and ablated baselines, this study offers evidence about what such designs can and cannot be expected to deliver on small institutional datasets.

2 Literature Review

A substantial body of work has examined how to recommend compatible learning partners in online and blended-learning environments. Hassan and Hamada (2016) proposed a recommender system that searches social-network sites to find and recommend learning peers based on users' posts, comments, and common friends, in order to enhance collaborative learning; the study, however, presented only the proposed methodology and did not report empirical results from an implementation. Bouchet et al. (2017) built a peer recommender within a MOOC that supplies each learner with a reachable list of peers willing to share knowledge, based on learner-profile similarity and behavioural actions; in the pre-MOOC phase, K-means clustering classified learners from most active to most passive using interaction logs of browsing, downloading, and viewing, and learners could message, favour, or ignore a recommended peer, with ignored peers no longer recommended, although the specific filtering technique used was not clearly specified.

Potts et al. (2018) introduced RIPPLE, an open-source platform for reciprocal peer recommendation that matches learners on mutually convenient time slots and complementary topics and roles, computing a reciprocal score that prioritises mutual benefit; the algorithm outperformed a non-reciprocal baseline on scalability, reciprocality, coverage, and quality, but relied on synthetic-data experiments without real-user validation. Chounta (2019) proposed using a 'grey area' operationalisation of the Zone of Proximal Development (ZPD) to decide when a student is above, below, or within their ZPD, using machine-learning cognitive models to track knowledge state and dynamically assign study groups accordingly; the paper, however, presented no empirical results demonstrating the system's operation within an actual MOOC.

Nguyen (2020) addressed the isolation associated with online learning by proposing a Learning Partner Recommender System (LPRS) that matches students on education level, learning style, academic interest, motivation, and personality traits, developed in two phases — a literature-informed identification of matching criteria followed by design, implementation, and evaluation of a reciprocal-recommendation proof of concept — though the focus group used to validate the LPRS involved only eight students. Elghomary and Bouzidi (2019) explored the role of trust in MOOCs through a dynamic Peer Recommendation System built on a trust-management system incorporating trust-metric selection, trust-network modelling, and a feedback mechanism; the approach improved collaboration and recommendation accuracy through real-time trust updates, but its reliance on learner-supplied feedback leaves it vulnerable to biased or inaccurate input, and it did not cover all contextual factors that influence trust.

Bencito et al. (2020) developed an Android-based Study Buddy Pairing application using an Agile development lifecycle, evaluated against Android development standards covering visual design, functionality, compatibility, performance, and security, and found to perform well across most criteria as an aid to studying and peer tutoring. Garg and Goel (2021) proposed a user-similarity-based peer-recommendation approach that uses learner scores and demographic attributes to identify potential partners for isolated learners, following a three-step process of detecting isolation, diagnosing its cause,

and generating peer recommendations; the approach balanced scalability and coverage well, though scalability was assessed only for a single target learner at a time, and successful interaction still depends on reciprocity between matched learners.

Oktavia and Sujarwo (2021) developed an interactive recommender system that draws on LinkedIn data and an ontological model to map learning topics, competencies, and job requirements for identifying suitable learning partners among professionals and educators, but the study offered limited theoretical grounding and did not address how the system would integrate into existing higher-education frameworks. Bourkoukou et al. (2017) combined a knowledge-level assessment tool with collaborative filtering to classify learners by profile and recommend learning objects matched to their assessed understanding, finding that most students considered the system beneficial, although reliance on multiple-choice assessment may understate deeper comprehension, and end-of-course satisfaction surveys are susceptible to response bias. Shao et al. (2022) proposed a learning-partner recommendation framework based on the evolution of fine-grained learning interest (LPRF-E), extracting time-varying interest tags via TextRank and clustering and computing interest similarity through cosine similarity of word embeddings, further refined using learners' social influence; evaluated on Chinese University MOOC and Douban Book data, the framework achieved roughly 50% improvements in precision and recall over baselines, though at considerable implementation complexity and computational cost. Riad et al. (2023) applied collaborative filtering together with constrained Pearson correlation, adjusted cosine measures, and K-nearest neighbour algorithms to adapt learning-object recommendations to learners' learning style and motivation score, validated on the MovieLens100K dataset; while the technique improved adaptation quality, its reliance on a movie-recommendation dataset limits the extent to which findings generalise to genuinely educational contexts.

Ma'ady (2024) applied a naïve Bayes text-classification algorithm with a sigmoid activation function to match research collaborators from academic articles collected via Google Scholar, based on research topic, location, and case study, achieving 88.4% accuracy under k-fold cross-validation, though the underlying database schema may become difficult to maintain at scale and naïve Bayes may not suit all higher-education contexts without further transformation of its output. Shou et al. (2023) proposed a weighted heterogeneous information network method that integrates multiple learner-related object types (students, teachers, videos, exercises, knowledge points) and automatically generated meta-paths, optimised through Bayesian Personalised Ranking, and validated on three real datasets (Online, DBLP, and Aminer) with improved precision and recall over existing methods, although the study did not incorporate student feedback on recommendation relevance and relied on a single feature-selection approach.

Most directly related to the present study, Shou et al. (2024) proposed the Knowledge-Learner Joint Recommendation (KLJRec) model, a joint recommendation algorithm for knowledge concepts and learning partners built on an improved Multi-gate Mixture-of-Experts (MMoE) architecture. A heterogeneous information network (HIN) was constructed from a MOOC platform, with meta-paths selected to extract human–computer interaction and student–student interaction information; the model builds a graph in which students are nodes and interactions are edges, learns student embeddings within a multi-task framework, computes similarity scores between students, and returns the Top-K most likely beneficial partners for a given student. KLJRec outperformed baseline and hybrid recommender systems in precision, recall, and diversity, but the authors acknowledged that it does not effectively handle new learners with limited data (the cold-start problem) and lacks mechanisms for fairness and bias reduction, resulting in the over-recommendation of already-popular learners.

2.1 Research Gap

The reviewed literature shows steady progress in recommending learning partners and collaborators, yet two challenges remain largely unresolved: reliance on interaction-based embeddings makes it difficult to serve recently enrolled learners with no interaction history, and the absence of fairness-aware mechanisms allows popularity bias to persist. Although KLJRec improved on earlier systems by jointly recommending knowledge concepts and learning partners, it did not resolve either problem. This study examines a hybrid extension of KLJRec-style graph recommendation with a Profile Encoder, contextual compatibility scoring and MMR re-ranking, and tests its claimed benefits with a protocol in which ground truth is independent of the recommender's own filters.

3 Methodology

3.1 Dataset

The study used a dataset of 206 learner records, privately generated and collected from a distance learning institution, together with 400 recorded learner-partnership edges denoting prior successful study collaborations between learners. Each learner record comprises a student identifier, academic major (Business, Computer Science, Engineering, Mathematics, or Physics), grade point average, semester level (1 to 8), preferred study time (Morning, Afternoon, or Evening), number of courses completed, and a five-dimensional skill profile. For model training and evaluation, the 400 learner-partnership edges were randomly split 80/20 into 320 training edges and 80 held-out test edges, with the Graph Neural Network trained on the training-edge subgraph and the held-out edges reserved for evaluation. In addition to interaction data, the dataset's contextual attributes — academic major, semester level, and preferred study time — are particularly useful for addressing the cold-start problem, since newly enrolled learners may lack sufficient interaction history for graph-based recommendation but still carry these attributes from the moment of enrolment. No personal demographic attributes such as age or gender were used; the recommendation process focused solely on academic and contextual learner characteristics.

3.2 Problem Formulation and Proposed System

The baseline system generates recommendations from GNN embeddings learned over learner interactions alone. The hybrid extension adds three components: a Profile Encoder that generates embeddings from contextual attributes for learners without graph edges (cold-start handling); a contextual-compatibility score that measures agreement on semester, major, and preferred study time; and Maximum Marginal Relevance (MMR) re-ranking to balance relevance and diversity. The original design applied the contextual component as an exact-match filter; because this was found to shrink the candidate pool to fewer than two learners on average, the study compares it with a soft alternative in which compatibility is a continuous score blended with embedding similarity, as suggested during peer review.

Cold-start handling is achieved through the Profile Encoder, which estimates representations for new learners from feature similarities in major, semester, and preferred time, and is invoked automatically whenever interaction data is missing for a target learner. Popularity bias — the disproportionate recommendation of already-popular or highly connected learners, which leads to over-exposure of active learners and under-exposure of new or isolated ones — is mitigated through MMR re-ranking, which selects candidates one at a time to maximise relevance to the target learner while minimising similarity to already-selected candidates, and optionally rewarding underrepresented candidates. The MMR score is computed as:

$$MMR(i) = \lambda \cdot Rel(i, u) - (1 - \lambda) \cdot \max_{j \in S} Sim(i, j)$$

where u is the target learner, i is the candidate under evaluation, S is the set of already-selected recommendations, and $\lambda \in [0,1]$ is the relevance–diversity trade-off parameter (Carbonell & Goldstein, 1998). Relevance is the final candidate score. In the soft variant, the final score is $\beta \cdot \text{sim}(u,i) + (1 - \beta) \cdot \text{ctx}(u,i)$, where sim is cosine similarity rescaled to [0,1] and ctx is the mean of three indicators (same semester, same preferred time, same major); in the original hard variant, candidates failing any of the three indicators were removed before MMR.

The Graph Neural Network learns node embeddings hi through neighbourhood aggregation in the style of GraphSAGE (Hamilton et al., 2017), updating each learner's representation from connected peers' features via message passing:

$$h_i^{(k+1)} = AGGREGATE(h_i^{(k)}, \{h_j^{(k)} : j \in N(i)\})$$

where N(i) denotes the neighbours of node i. The Profile Encoder separately encodes semester, major, and preferred time into a latent vector zprofile. The hybrid embedding used for recommendation is a weighted combination:

$$z_hybrid = z_GNN \text{ if learner has } \geq 1 \text{ training edge, otherwise } z_profile$$

That is, the model selects between graph-derived and profile-based embeddings depending on whether a learner has existing collaboration edges in the training graph; both encoders are trained with a margin-based triplet objective on training edges only. The overall procedure preprocesses learner features, constructs the training graph, trains the GNN and Profile Encoder, generates embeddings for each target learner, computes similarity and contextual scores against all candidates, applies MMR re-ranking, and returns the top-K recommended partners.

3.3 System Architecture

The system comprises five layers. The Input and Preprocessing Layer processes student profile data for the 206 learners (semester, major, preferred study time, GPA, courses completed, and five skill scores), encodes categorical attributes numerically, and converts partnership records into graph edges. The Graph Construction Layer applies a GraphSAGE-style

GNN to learn node embeddings from neighbour aggregation. The Hybrid Embedding Layer introduces the Profile Encoder for students with no training connections. The Context-Aware Scoring layer computes the contextual-compatibility score (exact-match filtering in the original design, soft scoring in the revised design). Finally, the Recommendation Layer applies MMR re-ranking and returns the Top-K recommended partners.

3.4 Recommendation Algorithm

The recommendation procedure, summarised in Algorithm 1 (see Appendix), begins by loading and preprocessing the learner dataset and encoding major, semester, and preferred study time. An interaction graph is then constructed with learners as nodes and recorded interactions as edges, with 80% of edges reserved for training and 20% held out for evaluation. A Graph Neural Network is trained on the training edges to learn graph-based embeddings, and a Profile Encoder is trained on the same edges (using profile features only, without graph access) to generate profile-based embeddings. For each target learner, the algorithm checks whether the learner has at least one training edge: if so, the GNN embedding is used; otherwise, the Profile Encoder embedding is used, which constitutes the system's cold-start handling mechanism. A final score is computed for each candidate as a weighted blend of embedding similarity and the soft contextual-compatibility score described above; Maximum Marginal Relevance re-ranking is then applied to this score to maximise relevance while improving diversity and reducing popularity bias, before the Top-K highest-ranked learners are returned as the final recommendation.

All models were implemented from scratch in Python with NumPy: GraphSAGE-style mean neighbour aggregation (with the learner's own features included) followed by a two-layer MLP encoder, manual backpropagation, and the Adam optimiser. Both encoders were trained on training edges only with a margin-based triplet objective (margin = 1.0) and per-epoch random negative sampling, 300 epochs, learning rate 0.02, hidden size 64, embedding size 32. A pairwise attraction-only objective without negatives was tried first and collapsed all embeddings (standard deviation across learners about 0.01), so it was not used. The 400 edges were randomly split 80/20 (320 training, 80 test), and the whole pipeline was repeated over five random splits and model initialisations; results are reported as mean ± standard deviation. Only learners with at least one held-out test edge were evaluated (about 108 per split, of whom about 5 had no training edge and form the cold-start subgroup), since Precision@K and Recall@K are undefined otherwise. Ground truth was the set of held-out partners. Compared methods were: random; popularity-only (training degree); GNN-only; Profile Encoder only; hybrid embedding (GNN or Profile embedding according to edge availability); hybrid with the original hard filter and MMR; hybrid with soft contextual score; and hybrid with soft score and MMR ($\beta = 0.7$, $\lambda = 0.7$, selected from a single sweep over $\beta \in \{0.3, 0.5, 0.6, 0.7, 0.9\}$ on one split; no separate validation set was used, so these are not tuned on held-out data).

3.5 Evaluation Metrics

Model performance was assessed using five metrics. Precision@K is the fraction of the K recommended partners that appear among a learner's held-out partners. Recall@K is the number of held-out partners retrieved divided by min(number of held-out partners, K). F1@K is the harmonic mean of Precision@K and Recall@K. NDCG@K rewards held-out partners ranked higher, computed as DCG@K / IDCG@K with binary relevance. The Popularity Bias Index (PBI) is the mean training-graph degree of the recommended learners, averaged over evaluated learners; lower values indicate less concentration on well-connected learners, and the population mean degree (about 3.1) corresponds to random recommendation.

4 Results and Discussion

Tables 1 to 3 report all models at K = 5, 10 and 15 over learners with at least one held-out test edge, as mean ± standard deviation across five random 80/20 edge splits. Table 4 reports the cold-start subgroup (learners with no training edge) at K = 10. The GNN-only model is the baseline; the original design is the hybrid with hard filter and MMR.

Table **1**

Evaluation Metrics at K = 5 (mean ± SD over 5 splits)

Model	Precision@K	Recall@K	F1@K	NDCG@K	PBI
Random	0.0059 ± 0.0037	0.0209 ± 0.0158	0.0088 ± 0.0059	0.0161 ± 0.0108	3.10 ± 0.05
Popularity-only	0.0082 ± 0.0040	0.0268 ± 0.0129	0.0119 ± 0.0055	0.0200 ± 0.0094	7.64 ± 0.41
GNN-only (baseline)	0.0048 ± 0.0029	0.0141 ± 0.0087	0.0068 ± 0.0041	0.0082 ± 0.0051	3.31 ± 0.05
Profile Encoder only	0.0107 ± 0.0044	0.0400 ± 0.0118	0.0163 ± 0.0063	0.0238 ± 0.0043	3.27 ± 0.05
Hybrid embedding (GNN/Profile switch)	0.0052 ± 0.0024	0.0146 ± 0.0079	0.0072 ± 0.0034	0.0090 ± 0.0048	3.28 ± 0.11
Hybrid + hard filter + MMR (original design)	0.0037 ± 0.0045	0.0029 ± 0.0037	0.0032 ± 0.0039	0.0034 ± 0.0042	2.28 ± 0.11
Hybrid + soft context score	0.0076 ± 0.0045	0.0217 ± 0.0156	0.0107 ± 0.0067	0.0145 ± 0.0111	3.29 ± 0.09
Hybrid + soft context score + MMR	0.0067 ± 0.0037	0.0203 ± 0.0137	0.0096 ± 0.0058	0.0131 ± 0.0093	2.97 ± 0.11

Table **2**

Evaluation Metrics at K = 10 (mean ± SD over 5 splits)

Model	Precision@K	Recall@K	F1@K	NDCG@K	PBI
Random	0.0067 ± 0.0013	0.0419 ± 0.0113	0.0112 ± 0.0022	0.0242 ± 0.0087	3.12 ± 0.07
Popularity-only	0.0080 ± 0.0015	0.0509 ± 0.0091	0.0132 ± 0.0025	0.0287 ± 0.0080	7.19 ± 0.28
GNN-only (baseline)	0.0068 ± 0.0014	0.0410 ± 0.0132	0.0113 ± 0.0025	0.0181 ± 0.0059	3.30 ± 0.06
Profile Encoder only	0.0095 ± 0.0046	0.0656 ± 0.0241	0.0161 ± 0.0076	0.0329 ± 0.0087	3.32 ± 0.04
Hybrid embedding (GNN/Profile switch)	0.0070 ± 0.0011	0.0429 ± 0.0104	0.0116 ± 0.0020	0.0192 ± 0.0048	3.32 ± 0.10

Hybrid + hard filter + MMR (original design)	0.0037 ± 0.0045	0.0029 ± 0.0037	0.0032 ± 0.0039	0.0034 ± 0.0042	2.28 ± 0.11
Hybrid + soft context score	0.0058 ± 0.0024	0.0315 ± 0.0144	0.0095 ± 0.0039	0.0184 ± 0.0105	3.29 ± 0.05
Hybrid + soft context score + MMR	0.0064 ± 0.0021	0.0401 ± 0.0125	0.0108 ± 0.0036	0.0200 ± 0.0088	2.98 ± 0.04

Table **3**

Evaluation Metrics at K = 15 (mean ± SD over 5 splits)

Model	Precision@K	Recall@K	F1@K	NDCG@K	PBI
Random	0.0058 ± 0.0020	0.0609 ± 0.0209	0.0104 ± 0.0035	0.0234 ± 0.0098	3.11 ± 0.02
Popularity-only	0.0092 ± 0.0024	0.0860 ± 0.0224	0.0161 ± 0.0041	0.0392 ± 0.0114	6.82 ± 0.24
GNN-only (baseline)	0.0064 ± 0.0021	0.0604 ± 0.0270	0.0112 ± 0.0039	0.0236 ± 0.0088	3.31 ± 0.07
Profile Encoder only	0.0089 ± 0.0036	0.0898 ± 0.0304	0.0159 ± 0.0063	0.0401 ± 0.0108	3.34 ± 0.04
Hybrid embedding (GNN/Profile switch)	0.0064 ± 0.0018	0.0600 ± 0.0243	0.0112 ± 0.0034	0.0242 ± 0.0074	3.33 ± 0.10
Hybrid + hard filter + MMR (original design)	0.0037 ± 0.0045	0.0029 ± 0.0037	0.0032 ± 0.0039	0.0034 ± 0.0042	2.28 ± 0.11
Hybrid + soft context score	0.0061 ± 0.0015	0.0561 ± 0.0146	0.0108 ± 0.0027	0.0254 ± 0.0099	3.28 ± 0.05
Hybrid + soft context score + MMR	0.0055 ± 0.0016	0.0546 ± 0.0150	0.0098 ± 0.0029	0.0241 ± 0.0092	2.97 ± 0.04

Three findings are consistent across K. First, the original hard contextual filter performed far worse than every other method. Second, replacing it with a soft contextual score, with or without MMR, restored performance to the level of the GNN-only baseline, with differences well inside one standard deviation. Third, no learned model reliably outperformed random or popularity-only recommendation.

4.1 Accuracy: Precision, Recall, F1 and NDCG

At K = 10, the hard-filter design reached Recall@10 of 0.0029 ± 0.0037, against 0.0410 ± 0.0132 for the GNN-only baseline, 0.0429 ± 0.0104 for the hybrid embedding, and 0.0401 ± 0.0125 for the soft score with MMR (Table 2). The hard filter left an average of fewer than two candidates per learner, so lists were far shorter than K and most held-out partners

were excluded by construction. The soft-score variants and the GNN-only baseline are not distinguishable at any K given the standard deviations (for example, NDCG@10 of 0.0200 ± 0.0088 versus 0.0181 ± 0.0059).

4.2 Trivial Baselines and Ablation

Random recommendation (Recall@10 = 0.0419 ± 0.0113, NDCG@10 = 0.0242 ± 0.0087) matched or exceeded the GNN-only baseline and the hybrid variants, and popularity-only recommendation was slightly higher still (Recall@10 = 0.0509 ± 0.0091, NDCG@10 = 0.0287 ± 0.0080). The Profile Encoder alone had the highest mean scores at K = 10 and K = 15 (for example, Recall@10 = 0.0656 ± 0.0241, NDCG@10 = 0.0329 ± 0.0087), but its intervals overlap with those of the popularity and random recommenders. The ablation therefore shows that neither the graph embeddings nor the contextual scoring added measurable accuracy over trivial recommenders on this dataset, and that the only component with a clear effect is the hard filter, which was harmful.

To check whether any model could in principle succeed, a signal diagnostic was run on the full 400-edge dataset. Connected learners were no more similar than randomly paired learners: mean absolute standardised differences were 1.06 to 1.21 for connected pairs versus 1.15 to 1.17 for random pairs on all eight numeric features, with only one of eight differences (skill_4) reaching $p < 0.05$ without correction for multiple comparisons. Connected pairs were slightly less likely than random pairs to share a major (0.177 versus 0.196), preferred time (0.310 versus 0.331) or semester (0.105 versus 0.125). Learner degree (mean 3.88, SD 2.07) was close to the SD expected from random (Poisson) edge formation (1.97), indicating no marked popularity structure. These checks indicate that the partnership links are statistically close to random with respect to the recorded attributes, which explains why no attribute-based or embedding-based model outperformed random recommendation.

4.3 Popularity Bias and Cold-Start Learners

The Popularity Bias Index is the mean training degree of recommended learners. Popularity-only recommendation had a PBI of 7.19 at K = 10, and random recommendation 3.12. The GNN-only baseline (3.30) and hybrid embedding (3.32) were at the random level. MMR re-ranking with the soft score reduced the PBI to 2.98 at K = 10 (2.97 at K = 5 and K = 15), and the hard-filter design reduced it further to 2.28 only because it returned very short lists. The MMR reduction is modest in absolute terms but is the only effect in the study that is separable from the random level in the intended direction; it was not accompanied by any accuracy cost relative to the GNN-only baseline.

Table **4**

Evaluation Metrics at K = 10, Cold-Start Learners Only (mean ± SD over 5 splits; about 5 learners per split)

Model	Precision@K	Recall@K	F1@K	NDCG@K	PBI
Random	0.0072 ± 0.0099	0.0722 ± 0.0988	0.0131 ± 0.0180	0.0367 ± 0.0466	3.18 ± 0.15
Popularity-only	0.0072 ± 0.0099	0.0199 ± 0.0257	0.0106 ± 0.0142	0.0121 ± 0.0166	7.20 ± 0.28
GNN-only (baseline)	0.0050 ± 0.0100	0.0125 ± 0.0250	0.0071 ± 0.0143	0.0056 ± 0.0113	3.23 ± 0.31
Profile Encoder only	0.0044 ± 0.0089	0.0333 ± 0.0667	0.0077 ± 0.0155	0.0172 ± 0.0344	3.31 ± 0.38
Hybrid embedding (GNN/Profile switch)	0.0100 ± 0.0200	0.0625 ± 0.1250	0.0162 ± 0.0325	0.0290 ± 0.0580	1.76 ± 0.65

Hybrid + hard filter + MMR (original design)	0.0250 ± 0.0500	0.0125 ± 0.0250	0.0167 ± 0.0333	0.0195 ± 0.0390	1.80 ± 0.41
Hybrid + soft context score	0.0100 ± 0.0200	0.0250 ± 0.0500	0.0143 ± 0.0286	0.0265 ± 0.0529	2.44 ± 0.31
Hybrid + soft context score + MMR	0.0100 ± 0.0200	0.0625 ± 0.1250	0.0162 ± 0.0325	0.0362 ± 0.0724	2.31 ± 0.22

The cold-start subgroup contained only about five learners per split, and the standard deviations in Table 4 are as large as or larger than the means (for example, Recall@10 for the hybrid embedding of 0.0625 ± 0.1250), so no model comparison for cold-start learners is supported by these data. In the earlier single-split analysis the original hard-filter design scored zero on every metric for cold-start learners; across five splits it returned non-empty but very short lists. The cold-start claim therefore remains untested at a scale sufficient to evaluate it.

4.4 Discussion in Relation to Prior Work

These results differ from the near-perfect precision reported for the hybrid design in the original dissertation, which arose from a ground truth that was a superset of the recommender's own filter, and from an embedding objective that collapsed. Once corrected, the hybrid design gives no measurable accuracy benefit on this dataset. A plausible explanation is data scale and signal: with 206 learners, 320 training edges and attributes limited to GPA, semester, courses completed, five skill scores, major and preferred time, the learned embeddings carry little information about who partners with whom, as shown by GNN-only recommendation performing at the level of random recommendation. This contrasts with studies such as Shou et al. (2024), which used much larger interaction graphs. The signal diagnostic reported above supports this explanation: connected learners were no more similar than random pairs on the recorded attributes, so no attribute-based method could be expected to succeed on this dataset.

The hard-filter failure is mechanistic rather than dataset-specific: requiring exact agreement on three categorical attributes leaves very few candidates, so any recommender using it will return short lists that exclude most true partners unless partners almost always share all three attributes. This is consistent with the use of profile-based similarity in earlier work such as Nguyen (2020) and Bourkoukou et al. (2017) as a soft similarity signal rather than a hard constraint. The popularity reduction from MMR parallels the diversity aims of Potts et al. (2018) and Garg and Goel (2021), but here it is a small effect. Overall, the study supports a conservative conclusion: contextual constraints should be soft, MMR can reduce popularity concentration without an accuracy penalty on this dataset, and accuracy claims for hybrid partner recommenders need non-circular evaluation and trivial baselines.

4.5 Justification of K Values

K = 5, 10, and 15 were selected at random to test how well the model will perform at different value of k and because they are widely used in recommender-system evaluation studies, allowing algorithm effectiveness to be assessed at different recommendation depths. K = 5 evaluates the quality of the most highly ranked recommendations that users are most likely to examine; K = 10 provides a balanced assessment; and K = 15 evaluates the model's ability to sustain recommendation relevance across a larger list. Using multiple values of K enables a more comprehensive evaluation and facilitates comparison with existing recommendation studies.

5 Conclusion and Recommendations

This study re-evaluated a hybrid extension of graph-based learning partner recommendation using held-out partnership edges as ground truth, five random splits, and trivial and ablated baselines. Two flaws in the original evaluation, a circular ground truth and a collapsed embedding objective, were identified and corrected. The original hard contextual filter was shown to be harmful, reducing Recall@10 from about 0.04 to 0.003; replacing it with a soft compatibility score removed

the harm. MMR re-ranking reduced the Popularity Bias Index (2.98 versus 3.30 for the GNN-only baseline at K = 10) without a measurable accuracy cost. However, no learned model reliably outperformed random or popularity-only recommendation, and the cold-start subgroup was too small for any conclusion.

The contributions are therefore: (i) identification and correction of a circular evaluation protocol and a collapsing embedding objective; (ii) evidence that exact-match contextual filtering is harmful and a soft-score alternative; and (iii) a negative result, with trivial baselines and an ablation, showing that the available data do not support accuracy gains from graph or profile embeddings, together with a small but measurable popularity-bias reduction from MMR. The accuracy and cold-start benefits claimed for the hybrid design are not supported by this dataset. Future work should evaluate the design on larger partnership graphs with richer behavioural features, include a validation split for tuning β and λ, test against strong graph baselines such as LightGCN or GraphSAGE implemented in a deep learning framework, and use a larger cold-start sample, possibly by simulating cold-start through removing all edges of selected learners.

6 Limitations

This study has several limitations. The dataset is small and sparse (206 learners, 400 edges), and its partnership links may have been formed independently of the recorded attributes, which limits what any method can learn. Results use five random splits, which gives only a rough estimate of variance, and β and λ were chosen from a single sweep on one split without a held-out validation set. The cold-start subgroup contained about five learners per split, so cold-start conclusions are not supported. The models were implemented from scratch in NumPy rather than with a deep learning framework, and stronger graph baselines (LightGCN, NGCF, PinSage) were not run. Results on this private dataset may not generalise to other institutions or larger platforms.

7 Declarations

Funding: This research received no external funding from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of interest: The authors declare that they have no conflict of interest related to the research presented in this manuscript.

Ethics statement: The dataset contains only anonymised learner identifiers, programme attributes and partnership links; it contains no names or contact details. We got permission from the institution to use the data from the institution to test run the experiments.

8 References

Bencito, G. R. L., Dabu, F. N., & Santiago, C. S., Jr. (2020). Development and evaluation of study-buddy pairing application for Android. *Journal of Environmental Science, Computer Science and Engineering & Technology*, *9*(2), 230–237. https://doi.org/10.24214/jecet.B.9.2.23037

Bouchet, F., Labarthe, H., Bachelet, R., & Yacef, K. (2017). Who wants to chat on a MOOC? Lessons from a peer recommender system. In Digital education: Out to the world and back to the campus (pp. 150–159). Springer. https://doi.org/10.1007/978-3-319-59044-8_17

Bourkoukou, O., El Bachari, E., & El Adnani, M. (2017). A recommender model in e-learning environment. *Arabian Journal for Science and Engineering*, *42*(2), 607–617. https://doi.org/10.1007/s13369-016-2292-2

Carbonell, J., & Goldstein, J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 335–336). ACM.

Chounta, I.-A. (2019). Using learning analytics to provide personalized recommendations for finding peers. arXiv. https://arxiv.org/abs/1910.07381

Elghomary, K., & Bouzidi, D. (2019). Dynamic peer recommendation system based on trust model for sustainable social tutoring in MOOCs. In 2019 International Conference on Smart Systems and Data Science (ICSSD). IEEE. https://doi.org/10.1109/ICSSD47982.2019.9003154

Garg, M., & Goel, A. (2021). A data-driven approach for peer recommendation to reduce dropouts in MOOC. In Advances in computing and network communications: Proceedings of CoCoNet 2020 (Lecture Notes in Electrical Engineering, Vol. 735, pp. 217–229). Springer. https://doi.org/10.1007/978-981-33-6977-1_18

Hamilton, W. L., Ying, R., & Leskovec, J. (2017). Inductive representation learning on large graphs. Advances in Neural Information Processing Systems, 30.

Hassan, M., & Hamada, M. (2016). Recommending learning peers for collaborative learning through social network sites. In 2016 7th International Conference on Intelligent Systems, Modelling and Simulation (ISMS) (pp. 60–63). IEEE. https://doi.org/10.1109/ISMS.2016.22

Ma'ady, M. N. P. (2024). Strengthening research partner collaboration in higher education for searching innovation through machine learning-based recommender system. *International Journal of Systematic Innovation*, *8*(1), 27–35. https://doi.org/10.6977/IJoSI.202403_8(1).0003

Nguyen, T. T. A. M. (2020). Learning Partner Recommender System (LPRS): Promoting informal online learning communities in higher education [Doctoral dissertation, Monash University]. Monash University Bridges. https://doi.org/10.26180/12486329

Oktavia, T., & Sujarwo, S. (2021). Interactive recommender system for identifying learning partner. *International Journal of Emerging Technology and Advanced Engineering*, 11(6), 78–86. https://doi.org/10.46338/ijetae0621_09

Potts, B. A., Khosravi, H., Reidsema, C., Bakharia, A., Belonogoff, M., & Fleming, M. (2018). Reciprocal peer recommendation for learning purposes. In Proceedings of the 8th International Conference on Learning Analytics and Knowledge (pp. 226–235). ACM. https://doi.org/10.1145/3170358.3170400

Riad, M., Qbadou, M., Aoula, E.-S., & Gouraguine, S. (2023). The new e-learning adaptation technique based on learner's learning style and motivation. *Journal of Education and Learning (EduLearn)*, *17*(3), 472–482. https://doi.org/10.11591/edulearn.v17i3.20826

Shao, M.-M., Jiang, W.-J., Wu, J., Shi, Y.-Q., Yum, T.-S., & Zhang, J. (2022). Improving friend recommendation for online learning with fine-grained evolving interest. *Journal of Computer Science and Technology*, *37*(6), 1444–1463. https://doi.org/10.1007/s11390-021-2124-z

Shou, Z., Chen, Y., Wen, H., Liu, J., Mo, J., & Zhang, H. (2024). Research on joint recommendation algorithm for knowledge concepts and learning partners based on improved multi-gate mixture-of-experts. *Electronics*, *13*(7), Article 1272. https://doi.org/10.3390/electronics13071272

Shou, Z., Shi, Z., Wen, H., Liu, J., & Zhang, H. (2023). Learning peer recommendation based on weighted heterogeneous information networks on online learning platforms. *Electronics*, *12*(9), Article 2051. https://doi.org/10.3390/electronics12092051

APPENDIX

Algorithm 1: Hybrid-KLJRec-Recommend(student_id, K, λ, α)

Input: Learner dataset D, interaction graph G(V,E), learner features F, target learner u, number of recommendations K.

Output: Top-K learning partner recommendations.

1. Load learner dataset D.
2. Preprocess learner features: encode major, semester, and preferred study time.
3. Construct interaction graph G(V,E): represent learners as nodes and learner interactions as edges.
4. Train the Graph Neural Network (GNN) to learn graph-based learner embeddings.
5. Train the Profile Encoder to generate profile-based embeddings from learner attributes.
6. For each target learner u: if u has interaction history, generate the embedding using the GNN; otherwise generate the embedding using the Profile Encoder (cold-start handling).
7. Compute similarity scores between learner u and all candidate learners.

8. Rank candidates according to similarity scores.
9. Compute a contextual-compatibility score for each candidate (agreement on major, semester, preferred study time); the original design instead removed candidates that did not match all three.
10. Blend the contextual-compatibility score with embedding similarity, then apply Maximum Marginal Relevance (MMR) re-ranking to the blended score to maximise relevance, improve diversity, and reduce popularity bias.
11. Select the Top-K highest-ranked learners.
12. Return the Top-K learning partner recommendations.

The recommendation function was implemented in Python using PyTorch and PyTorch Geometric. For a target student, the function extracts the student's semester, preferred time, and major; computes graph embeddings via the trained GNN and profile embeddings via the trained Profile Encoder; selects the graph embedding where the student has existing edges and the profile embedding otherwise; computes cosine similarity against all other students; filters candidates to those sharing the same semester, preferred time, and major; and iteratively selects the Top-K candidates using an MMR score that combines relevance, a diversity penalty based on similarity to already-selected candidates, and a minor fairness weight that favours less-connected learners.

Publisher's Note

The views, interpretations, and conclusions expressed in this article are solely those of the author(s) and do not necessarily reflect the views or policies of the Journal of Studies in Science and Mathematics Education (JOSSME), its publisher, editorial board, editors, reviewers, or the authors' affiliated institutions. The publisher remains neutral with respect to jurisdictional claims in published maps, institutional affiliations, and other representations. Any products, materials, methods, or manufacturers mentioned or evaluated in this article are presented solely for scholarly purposes and do not constitute endorsement, recommendation, or guarantee by the publisher.